Digital Cultural Heritage and the Crowd

Trevor Owens

Trevor Owens (trevor.johnowens@gmail.com), digital archivist, the Library of Congress, Washington, D.C.

Abstract
Libraries, archives, and museums have a long history of collaboration with members of the public. There is already considerable interest in extending this relationship, inviting members of the public, often referred to as "the crowd," to tag and classify, transcribe, organize, and otherwise add value to digital cultural heritage collection content. In this essay I connect current discussions of crowdsourcing with the mission and values of cultural heritage organizations and offer a framework for thinking about distinct components of different kinds of projects that have been lumped together.

The Two Problems with Crowdsourcing: Crowd and Sourcing

Libraries, archives, and museums have a long history of collaboration with members of the public. There is already considerable interest in extending this relationship, inviting members of the public, often referred to as "the crowd," to tag and classify, transcribe, organize, and otherwise add value to digital cultural heritage collection content.[1] In this essay I connect current discussions of crowdsourcing with the mission and values of cultural heritage organizations and offer a framework for thinking about distinct components of different kinds of projects that have been lumped together.

There are two primary problems with bringing the idea of crowdsourcing into cultural heritage organizations. Both the idea of the crowd and the notion of sourcing are problematic terms. The most successful crowdsourcing projects in libraries, archives, and museums have not involved massive crowds and they have very little to do with outsourcing labor.

The term "crowd" is somewhat misleading, since most successful crowdsourcing projects do not rely on large, anonymous masses of people. These projects succeed by inviting participation from engaged members of the public. The success is built upon a long-standing tradition of volunteerism and involvement of citizens in the creation and development of public good.

For example, the New York Public Library's menu transcription project, "What's on the Menu?" invites members of the public to help transcribe the names and costs of menu items from digitized copies of menus from New York restaurants. Any interested person is free to visit the project website and start transcribing menus. But in practice, a dedicated community of foodies, New York history buffs, chefs, and self-motivated individuals are the ones who are excited about offering their time and energy to help volunteer and contribute, improving the public library's resource for others to use (see Vershbow 2011).[2] (Years ago, NYPL volunteers sat at a desk in the reading room

cataloging the original collection of menus.[3]) The technology enabling this participation may be relatively new. However, the menu project is a continuation of a longstanding tradition of inviting members of the public to help refine, enhance, and support resources. In short, "crowdsourcing" the menu transcription is not about crowds at all. It is about using digital tools to invite in volunteers in much the same way that members of the public have volunteered to help organize and add value to the NYPL collection in the past.

The problem with the term "sourcing" is its association with labor. Wikipedia's definition of crowdsourcing helps further clarify this relationship: "Crowdsourcing is a process that involves outsourcing tasks to a distributed group of people." The keyword in that definition is "outsourcing." Crowdsourcing is a concept that was invented and defined in the business world and it is important that we reconsider the things that change when we bring the term into cultural heritage.

At this point, we need to think for a moment about what we mean by terms like work and labor. While it might be acceptable for commercial entities to coax individuals to provide free labor, the ethical implications of such methods should give pause to cultural heritage organizations. It is critical here to unpack some of the different meanings we ascribe to "work." When we use the term "a day's work" we are directly referring to labor—to the kinds of work that one engages in as a financial transaction for pay. In contrast, when we use the term "work" to refer to someone's "life's work," we are referring to something that is significantly different. The former is about acquiring the resources one needs to survive. The latter is about the activities that we engage in that give our lives meaning. In cultural heritage, we have values and missions and we are in an opportune position to invite the public to participate. However, when we do so, we should not treat them as a crowd, and we should not attempt to source labor from them. When we invite the public we should do so under a different set of terms.

Citizen Scientists, Archivists, and the Meaning of Amateur

Some of the projects that fit under the heading of crowdsourcing have chosen very different terms to describe themselves. The names of these projects highlight the extent to which they invite participation from members of the public who identify with particular professional occupations and their characteristic ways of thinking. For example, the "Galaxy Zoo" project, which invites users to identify different types of galaxies in images collected from the Sloan Digital Sky Survey, refers to its users as citizen scientists.[4]

Similarly, the United States National Archives and Records Administration recently launched a crowdsourcing project, the Citizen Archivists Dashboard; its terminology invites "citizens"—not some anonymous "crowd"—to participate.[5] While these citizen archivists and scientists are not professional in the literal sense (they are unpaid for their expertise), they connect with something a bit different than volunteerism. They are amateurs in the truest and best possible sense of the term.

Amateurs have a long and vibrant history as contributors to the public good. Coming to English from French, the term *amateur* means a "lover of." The primarily negative connotations we place on the term are a relatively recent development. In other eras, the term "amateur" simply meant that someone was not a professional, that is, they were not paid for these particular labors of love. Charles Darwin, Gregor Mendel, and many others who made significant contributions to the sciences did so as amateurs. As a continuation of this line of thinking, the various Galaxy Zoo projects see the amateurs who participate

as peers, in many cases listing them as co-authors of academic papers published as a result of their work. I suggest that we think of crowdsourcing not as extracting labor from a crowd, but as a way for us to invite the participation of amateurs (in the non-derogatory sense of the word)—those with every bit of the potential to be a Darwin or Mendel—in the creation, development, and further refinement of public good.

Toward a better, more nuanced, notion of Crowdsourcing

Fighting against a word, however, is rarely a successful project. From here on I will continue to use and refine a definition for crowdsourcing that works for the cultural heritage sector. I will explain what I think are the four key components of this ethical crowdsourcing, which invites members of the public to participate as amateurs in the production, development, and refinement of public goods. These fall into the following four considerations: human computation; the wisdom of crowds; thinking of tools and software as scaffolding; and the psychology of participant motivation. Each of these phrases suggests a series of questions to ask of any cultural heritage crowdsourcing project.

I believe these four concepts provide us with the descriptive language to understand what makes crowdsourcing such a powerful tool—not only for improving and enhancing data related to cultural heritage collections, but also as a way for deep engagement with the public.

Human Computation

"Human computation" is grounded in the fact that human beings are able to process particular kinds of information and make judgments in ways that computers can't. To this end, there are a range of projects described as crowdsourcing that are anchored in the idea of treating people as processors. The best way to explain the concept is through a few examples of the role that human computation plays in crowdsourcing.

ReCaptcha is a great instance of how the processing power of humans can be harnessed to improve cultural heritage collection data (see Ahn, et al. 2008).[6] Most readers will be familiar with the little ReCaptcha boxes we fill out when we need to prove that we are in fact a person and not an automated system attempting to log into some site. Our ability to read the strange and messed up text in those little boxes proves that we are people. But in the case of ReCaptcha it also helps us correct the relatively poor quality of the automated text of digitized *New York Times* and Google Books. The same capability that allows people to differentiate themselves from machines is what allows us to help improve the full text search of the digitized *New York Times* and Google Books collections.

The principles of human computation are similarly on display in the Google Image Labeler game. From 2006-2011 the Google image labeler game invited members of the public to describe and classify the content of images in a head to head game.[7] For example, imagine that a player is viewing a red car. Somewhere else in the world, another player is also viewing that image. Each player is invited to key in labels for the image, with a series of "off-limits" words which have already been associated with the image (perhaps red and car). Each label I can enter which matches a label entered by the other player results in game points (perhaps entering road, identifying the road underneath the car in the picture). The game has inspired an open source version specifically designed

for use at cultural heritage organizations.[8] The design of this interaction is such that, in most cases, it results in generating increasingly high quality and detailed descriptions of images.

Both the image labeler and ReCaptcha are fundamentally about tapping into the capabilities of people to process information. I had earlier suggested that the kind of crowdsourcing I want us to be thinking about is not about labor, yet these kinds of human computation projects are often fundamentally about labor. This is most clearly visible in Amazon's Mechanical Turk, a site that allows anyone to program and post rudimentary tasks (transcribing text, write short reviews, identifying if images are obscene) and pay people (often pennies per individual task) to complete them.

Mechanical Turk's tagline is that it "gives businesses and developers access to an on-demand, scalable workforce" where "workers select from thousands of tasks and work whenever it's convenient." The labor focus of this site should give pause to those in the cultural heritage sector, particularly those working for public institutions. There are legitimate concerns that this kind of labor could be serving as a kind of "digital sweatshop."[9]

While there are reasons to worry about the potentially exploitive properties of projects like Mechanical Turk, it is important to realize that many of the same human computation activities which one could run through Mechanical Turk, like describing images or transcribing text, are not really the same kind of labor when they are situated as projects of "citizen science."

For example, the Galaxy Zoo invites individuals to identify galaxies. The activity is fundamentally similar to the Google image labeler game. Users are presented with an image of a galaxy and invited to classify it based on a simple set of taxonomic information. While the interaction is more or less the same, the change in context is essential.

Galaxy Zoo invites amateur astronomers to help classify images of galaxies. While the image identification task here is more or less the same as the others previously discussed, this site—at least in the early stages of the project—often gave these amateur astronomers their first-ever opportunity to see these stellar objects.[10] These images were all captured by a robotic telescope. The first Galaxy Zoo participants who looked at these images were actually the first people ever to see each of these stellar objects. In this case, the amateurs who catalogue these galaxies do so because they want to contribute to science (see Raddick, et al., 2010). Beyond engaging in this classification activity, the Galaxy Zoo project also invites members to discuss the galaxies in a forum. This discussion forum ends up representing a very different kind of crowdsourcing, one based not so much on the idea of human computation but instead on a notion which I refer to here as the wisdom of crowds.

A Key question emerges from the concept of human computation: How could we use human judgment to augment computer-processable information? It would be a waste of the public's time to invite people to complete a task that a computer could do. The value that human computation offers is the question of how the unique capabilities of people can be integrated into systems for the creation of public goods.

The Wisdom of Crowds, or Why Wasn't I Consulted

The phrase comes from James Surowiecki's book, *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business,*

Economies, Societies and Nations (2004). Surowiecki talks about a range of examples of how crowds of people can create important and valuable kinds of knowledge. Unlike human computation, the wisdom of crowds is not about highly structured activities. In Surowiecki's argument, the wisdom of crowds is an emergent phenomena resulting from how discussion and interaction platforms like wikis enable individuals to add to and edit each other's work.

The "wisdom of crowds" notion tends to come with a bit too much utopian baggage. It is hard to believe that a technology like the Web would substantively change our rather imperfect human nature. In light of this, Paul Ford's reformulation of the wisdom of crowds is particularly valuable. Ford, in his blog,[11] suggests that the heart of this matter is that the Web, unlike other mediums, is particularly well suited to answer the question "Why wasn't I consulted?"

> Why wasn't I consulted, which I abbreviate as WWIC, is the fundamental question of the Web. It is the rule from which other rules are derived. Humans have a fundamental need to be consulted, engaged, to exercise their knowledge (and thus power), and no other medium that came before has been able to tap into that as effectively.

He goes on to explain a series of projects that succeed because of their ability to energize this human desire to be consulted.

> If you tap into the human need to be consulted you can get some interesting reactions. Here are a few: Wikipedia, StackOverflow, Hunch, Reddit, MetaFilter, YouTube, Twitter, StumbleUpon, About, Quora, Ebay,Yelp, Flickr, IMDB, Amazon.com, Craigslist, GitHub, SourceForge, every messageboard or site with comments, 4Chan, Encyclopedia Dramatica. Plus the entire Open Source movement.

Each of these cases taps into our desire to respond. The comments sections in online news articles, or our ability to sign-up for an account and start providing our thoughts and ideas on Twitter or in a Tumblr, are fundamentally about this desire to be consulted.

Returning to the example from Galaxy Zoo, the carefully designed human computation classification exercise provides one kind of input, while the project's active Web forums capitalize on the opportunity to consult. Importantly, some of the most valuable discoveries in the Galaxy Zoo project—including an entirely new kind of green-colored galaxy—were the result of users sharing and discussing some of the images from the classification exercise in the open discussion forums.

A key question emerges from the concept of the "Wisdom of Crowds": How could we empower and consult with a community of users? Unlike human computation, the goal here is not users' ability to process information or make judgments, but rather their desire to provide their opinion.

Tools as Scaffolding

Helping someone succeed is often about getting them the right tools. Consider the function of scaffolding. Whether in constructing a house or building a workforce, scaffolding puts workers in a position to do their job. By standing on the scaffolding, they

are able to do their work without thinking about what supports them. In the activity of the work, the tool disappears and allows them to go about their tasks while taking for granted that they are suspended six or seven feet in the air. It is fruitful to think about a wide range of tools that serve as scaffolds.[12]

All tools can act as scaffolds to enable us to accomplish a particular task. This observation has a direct translation into the design of online tools as well. For example, before joining the Library of Congress, I worked on the Zotero project, a widely used free and open-source reference management tool that helps researchers collect, organize and cite bibliographic information. Zotero was translated into more than 30 languages by its users. The translation process was made significantly easier through BabelZilla, an online community for developers and translators of extensions for Firefox web browser. BabelZilla has a robust community of users who work to create translations. One of the neatest features of this platform is that texts that need to be translated from the source code are stripped out so that potential translators are presented with a simple Web page where they just type in translations of the lines of text.

This not only makes the process much simpler and quicker, it also means that potential translators need absolutely no knowledge of the programming in order to contribute to a translated version. Without BabelZilla, a potential translator would need to know about how Firefox works in a technical sense, and be comfortable with editing files in technical formats like XML in a text editor. But BabelZilla scaffolds users over that required knowledge and just lets them fill out translations in a simple text box on a Web page.

If we again return to the Galaxy Zoo example, we can now think of the classification game as a scaffold that allows interested amateurs to participate at the cutting edge of scientific inquiry. In this scenario, the entire technical apparatus, all of the technical equipment used in the Sloan Digital Sky Survey, the design of the Galaxy Zoo site, and all the work of the scientists and engineers that went into those systems are part of one big scaffold that puts users in a position to contribute to the frontiers of science through their actions on the website, without needing the skills and background of a professional scientist.[13]

Additional key question emerge from the notion of scaffolding: How can our tools act as scaffolds to help make the most of user efforts? What expertise can we embed inside the design of our tools to magnify user efforts? How can our tools put a potential user in exactly the right position, with the right knowledge, just at the moment he or she needs it, to accomplish a given activity?

Understanding Participant Motivation

Ben Brumfield runs a range of crowdsourcing transcription projects.[14] At one point he noticed that one of his power users was slowing down—cutting back significantly on the time spent transcribing these manuscripts. The user explained that there weren't many manuscripts left to transcribe. For this user, the two or three hours a day spent working on transcriptions were important. Participating in this project was so satisfying, and contributing to it involved so much self-image, that the user needed to ration out those remaining pages to make sure that the experience lasted as long as it could. When Brumfield found that out, he quickly put up some more pages. This particular story illustrates several points about what motivates us.

After people's basic needs are covered, they tend to be primarily motivated by things that are not financial. People identify and support causes and projects that provide them with a sense of purpose. They establish and sustain their identity and sense of self through their actions. People get meaning from doing things that matter to them. They find a sense of belonging by being a part of something bigger than themselves.[15] Projects that can tap into these identities and purposes while providing meaning to people's lives are projects that—far from exploiting people—can provide a way for them to connect with each other and make meaningful contributions to the public good.

This is one of the places where libraries, archives, and museums have the most to offer. As stewards of cultural memory, our institutions have a strong sense of purpose and their explicit mission is to serve the public good. This notion of motivation prompts further key questions for projects: Whose sense of purpose does this project connect to? What identities are involved? What kinds of people does this project matter to? And how can we connect with and invite the participation of those people?

Why are we putting cultural heritage collections online again?

There are many reasons that we put digital collections online. The single most important is to make history accessible so that we can invite students, researchers, teachers, and the public to explore and connect with our past. Historians, librarians, archivists, and curators who share digital collections and exhibits can measure their success in moving toward this goal by how people use, reuse, explore, and understand these objects.

In general, these crowdsourcing projects are described as a means by which we can get better data to help enable the kinds of use and reuse that we want people to make of our collections. In this respect, the general idea of crowdsourcing is described as an instrument for getting people to help us with data that can make collections more accessible. Crowdsourcing does this—and more. In the process of developing these crowdsourcing projects we have stumbled onto something far more exciting than speeding up or lowering the costs of document transcription. An example will help illustrate.

Increased Use, Deeper Use: Crowdsourcing Civil War Diaries
Last year, the University of Iowa libraries crowdsourced the transcription of a set of Civil War diaries. According to Nicole Saylor, the head of Digital Library Services, the project was very successful (Owens 2011). The library got great transcriptions and ended up attracting more donors to support its work.

The project also succeeded in dramatically increasing visitors to the library's website. As Saylor explained, "On June 9, 2011, we went from about 1,000 daily hits to our digital library on a really good day, to more than 70,000." As great as all this is, as far as I'm concerned, something even more valuable also happened. When people came to transcribe the diaries, they engaged with the objects more deeply than they would have if transcription were not an option. Consider this quote from Saylor explaining how one particular transcriptionist interacted with the collection:

> The transcriptionists actually follow the story told in these manuscripts
> and often become invested in the story or motivated by the thought of
> furthering research by making these written texts accessible. One of our
> most engaged transcribers, a man from the north of England, has written

us to say that the people in the diaries have become almost an extended part of his family. He gets caught up in their lives, and even mourns their deaths. He has enlisted one of his friends, who has a Ph.D. in military history, to look for errors in the transcriptions already submitted. "You can do it when you want as long as you want, and you are, literally, making history," he once wrote us. That kind of patron passion for a manuscript collection is a dream. Of the user feedback we've received, a few of my other favorites are: "This is one of the COOLEST and most historically interesting things I have seen since I first saw a dinosaur fossil and realized how big they actually were." "I got hooked and did about 20. It's getting easier the longer I transcribe for him because I'm understanding his handwriting and syntax better." "Best thing ever. Will be my new guilty pleasure that I don't even need to feel that guilty about."

The transcriptions are great—they make the content more accessible—but as Saylor explains, "The connections we've made with users and their sustained interest in the collection is the most exciting and gratifying part." This is exactly as it should be! The invitation of crowdsourcing and the event of the project are the most valuable and precious user experiences that a cultural heritage institution can offer. It is essential that the project provide meaningful work. These projects invite the public to leave a mark and help enhance the collections. If the goal is to get people to engage with collections and with the past, then the transcripts are actually a wonderful byproduct of offering meaningful activities for the public to engage in.

Meaningful Activity is the Apex of User Experience for Cultural Heritage Collections

What crowdsourcing does (and most digital collection platforms fail to do) is to offer an opportunity for someone to do something more than consume information. When done well, crowdsourcing offers us an opportunity to provide meaningful ways for individuals to engage with and contribute to public memory. Not limited to being an instrument which enables us to better deliver content to end users, crowdsourcing is the best way to actually engage our users in the fundamental reason that these digital collections exist in the first place.

When we adopt this mindset, the money spent on crowdsourcing projects—in terms of designing and building systems, staff time to manage, and so on—cannot be compared to the costs of having someone transcribe documents on Mechanical Turk. The transcription of those documents is actually a precious resource, a precious bit of activity that could mean the world to someone. You can't ask users to take on just any task or obstacle—for example, if you asked users to transcribe documents that could easily be transcribed by computers, the whole project would lose its meaning and purpose. It isn't about Sisyphean tasks; it is about providing meaningful ways for the public to enhance collections while more deeply engaging and exploring them.

Just as Brumfield's user rationed out the transcription of those documents, we might actually think about crowdsourcing as one of the most precious experiences we can offer our users. Instead of simply giving them the ability to browse or poke around in digital collections, we can invite them to participate. We are in a position to let the users

of these collections leave a mark on the collections. Instead of browsing through a collection they literally become authors of our historical record.

Acknowledgment

Notes
1. For projects focused on classification, see Trant (2008). For a discussion of crowdsourcing in cultural heritage, see Holley (2010) and Smith-Yoshimura (2012).
2. Vershbow slides are online at http://www.digitalpreservation.gov/meetings/documents/ndiipp11/vershbow.pdf.
3. See Taranto (2011). Video available online at http://vimeo.com/38196574.
4. Galaxy Zoo accessed at http://www.galaxyzoo.org/.
5. Citizen Archivist Dashboard accessed at http://www.archives.gov/citizen-archivist/.
6. For some further context on both ReCaptcha and Google Image Labler, see Luis von Ahn, *Human Computation*, Google TechTalks, 2006, accessed at http://video.google.com/videoplay?docid=-8246463980976635143.
7. See Metadata Games, http://www.tiltfactor.org/metadata-games.
8. For some brief coverage of these discussions see Williams (2010a; 2010b).
9. For an account of the history of the Galaxy Zoo Project, see *Chris Lintott on The Galaxy Zoo*, 2010, accessed at http://www.youtube.com/watch?v=j_zQIQRr1Bo.
10. See Paul Ford, The Web is a customer service medium, Jan. 6, 2011, accessed at http://www.ftrain.com/wwic.html .
11. Here I am drawing on the Vygotskyan tradition of "scaffolding." See Wood, Bruner, and Ross (1976). On the idea of cultural mediation, see Vygotsky (1978).
12. This broader understanding of tools is best explained in Andy Clark's notion of cognitive extension (2008).
13. For background on Ben Brumfield's crowdsourcing work and projects, see his blog http://manuscripttranscription.blogspot.com/.
14. For a popular account of current thinking on the psychology of motivation, see Pink (2009). For more substantive but still accessible academic research on the subject, see essays in Elliot and Dweck, eds. (2005).

References

———. 2010b. The ethics of Amazon's Mechanical Turk. *The Chronicle of Higher Education*. *ProfHacker*, March 1, accessed at http://chronicle.com/blogs/profhacker/the-ethics-of-amazons-mechanical-turk/23010.

Ahn, Luis von, Ben Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-based character recognition via Web security measures. *Science* 321(5895): 1465–1468, doi:10.1126/science.1160379.

Clark, Andy. 2008. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.

Elliot, Andrew and Carol Dweck, eds. 2005. *Handbook of Competence and Motivation*. New York: Guilford Press.

Holley, Rose. 2010. Crowdsourcing: How and why should libraries do it? *D-Lib Magazine* 16(3/4) (March), accessed at http://www.dlib.org/dlib/march10/holley/03holley.html.

Owens, Trevor. 2011. Crowdsourcing the Civil War: Insights interview with Nicole Saylor. *The Signal: The Library of Congress Digital Preservation Blog*, Dec. 6. Accessed at http://blogs.loc.gov/digitalpreservation/2011/12/crowdsourcing-the-civil-war-insights-interview-with-nicole-saylor/.

Pink, Daniel. 2009. *Drive: The Surprising Truth About What Motivates Us*. New York: Riverhead.

Raddick, Jordan, Georgia Bracey, Pamela Gay, Chris Lintott, Phil Murray, Kevin Schawinski, Alexander Szalay, and Jan Vandenberg. 2010. Galaxy Zoo: Exploring the motivations of citizen science volunteers. *Astronomy Education Review* 9(1).

Smith-Yoshimura, Karen. 2012. *Social Metadata for Libraries, Archives, and Museums: Executive Summary*. Dublin, Ohio: OCLC Research.

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. New York: Doubleday.

Taranto, Barbara. 2011. Crowdsourcing metadata. Presented at the Coalition for Networked Information Fall 2011 Membership Meeting, Dec. 12-13.

Trant, Jennifer. 2008. Curating collections knowledge: Museums on the cyberinfrastructure. In *Museum Informatics: People, Information and Technology in Museums*, Paul F. Marty and Katherine Burton Jones, eds., 275-291. New York and Oxon: Routledge, 2008.

Vershbow, Ben. 2011. Bringing in the crowd: Effects, affects and a few (minor) defects. Paper presented at Make It Work: Improvisations on the Stewardship of Digital Information, conference of the National Digital Stewardship Alliance, Washington, D.C., July 19-21.

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*, Michael Cole, Vera John-Steiner, and Sylvia Scribner, eds. Cambridge: Harvard University Press.

Williams, George. 2010a. The reliability, efficiency, and affordability of Amazon's Mechanical Turk. *The Chronicle of Higher Education*. *ProfHacker*, Feb. 22, accessed at http://chronicle.com/blogs/profhacker/the-reliability-efficiencyaffordability-of-amazons-mechanical-turk/22994.

Wood, David, Jerome Bruner, Gail Ross. 1976.The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* 17: 89-100.