Putting the world's cultural heritage online with crowdsourcing

by
Frederick Zarndt
frederick@frederickzarndt.com
Chair, IFLA Newspapers Section
Consultant to Content Conversion Specialists, Digital Divide Data, and DL Consulting

Abstract

Following the splash made by National Library of Australia's Trove crowdsourced newspaper OCR text correction and tagging, more and more cultural heritage organizations have begun to use crowdsourcing for projects that would otherwise have been expensive (transcription of manuscripts and records, correction of OCR text to high accuracy) or computationally impossible (tagging images and articles with noisy text). Trove was not the first use of crowdsourcing for cultural heritage content -- that distinction belongs to Project Gutenberg / Distributed Proofreaders -- but Trove was the first to use crowdsourcing for a mass digitization project. In this paper we briefly examine crowdsourcing, a few cultural heritage projects using crowdsourcing, its economics, and motivations of the crowd.

1. Crowdsourcing overview

In recent years crowdsourcing has exploded. The word itself is much in vogue. On July 5, 2012, Wikipedia listed 101 references to other Wikipedia pages in its "Crowdsourcing" category whereas on January 22, 2010[1], it listed only 41 pages. Similarly on July 5, 2012, the Wikipedia page on "Crowdsourcing" itself lists 53 external references; on July 5, 2010, only 10 external references.

The word "crowdsourcing' was coined by Jeff Howe in the article "*The rise of crowdsourcing*" written for *Wired* magazine[2] in June 2006. In it Howe drafted 5 principles describing the *new labor pool*

1. The crowd is dispersed
2. The crowd has a short attention span
3. The crowd is full of specialists
4. The crowd produces mostly crap
5. The crowd finds the best stuff

[1] January 22, 2010, was the first capture of http://en.wikipedia.org/wiki/Category:Crowdsourcing made by the Internet Archive's Wayback Machine.

[2] Jeff Howe. "The rise of crowdsourcing." *Wired,* Issue 14.06, June 2006.

As we will see later, some of these principles apply to cultural heritage crowdsourcing (1, 5), others definitely do not (2, 3, 4).

Interestingly, Howe does not mention James Surowiecki’s 2004 book “*The wisdom of crowds.*”[3]. In his book Surowiecki describes several experiments and events where the aggregate task performance of a crowd of “normal” people with no special training is as good as or better than a single expert. Although apparently no one had written about this phenomenon prior to Surowiecki, the phenomenon itself is not too surprising: “two heads are better than one”, “dos cabezas piensan mejor que una”, “vier Augen sehen mehr als zwei”, “deux avis valent mieux qu’un”, “yhteistyö on voimaa”, and “三个臭皮匠，胜过诸葛亮” is a notion that is common to many cultures and languages.

What is crowdsourcing? According to Wikipedia[4]

> Crowdsourcing is a process that involves outsourcing tasks to a distributed group of people. ... the difference between crowdsourcing and ordinary outsourcing is that a task or problem is outsourced to an undefined public rather than a specific body, such as paid employees.

Or for those who prefer a more formal and pedantic definition, Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara surveyed crowdsourcing literature and research to develop this definition[5]

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.

[3] James Surowiecki. *The wisdom of crowds*. New York: Random House. 2004.

[4] Wikipedia contributors, "Crowdsourcing," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Crowdsourcing (accessed June 1, 2012).

[5] [1]Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. *Towards an integrated crowdsourcing definition*. Journal of Information Science XX(X). 2012. pp. 1-14.

On January 25, 2010, Wikipedia listed 34 crowdsourcing projects[6]. In July 2012, that list had grown to ~122 projects[7]. And of these 122 projects, 5 are connected with digitized books, journals, manuscripts, or records:

- National Library of Australia's Australian Historic Newspapers (http://trove.nla.gov.au/)

- Family Search Indexing (https://familysearch.org/volunteer/indexing)
- Transcribe Bentham (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham)

- Liljenquist Family American Civil War photographs donated to Library of Congress and crowdsourced on Flickr (http://www.flickr.com/photos/library_of_congress/sets/72157625520211184/)

- Distributed Proofreaders / Project Gutenberg (http://www.pgdp.net)
-
- California Digital Newspaper Collection (http://cdnc.ucr.edu/) uses crowdsourcing to correct OCR text in newspapers.

2. Crowdsourcing types

As one might guess from Wikipedia's long list of 122 projects, crowdsourcing projects are quite diverse. Wikipedia types crowdsourcing projects into one of 7 different categories[8] (1) crowdvoting, (2) wisdom of the crowd projects, (3) crowdfunding, (4) crowdpurchasing, (5) microwork, (6) competitions, and (7) implicit crowdsourcing. In his doctoral dissertation, Daren Brabham[9] classifies crowdsourcing projects as either one of (1) knowledge discovery and management, (2) broadcast search, (3) peer-vetted creative production, or (4) distributed human intelligence tasking. Enrique Estellés offers yet another typology[10]: (1) crowdcasting, (2) crowdcollaboration (crowdstorming, crowdsupport), (3) crowdcontent (crowdproduction, crowdsearching, crowdanalyzing), (4) crowdfunding, and (5) crowdopinion.

[6] January 25, 2010, was the first capture of http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects made by the Internet Archive's Wayback Machine.

[7] Wikipedia contributors, "List of crowdsourcing projects," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects (accessed July 5, 2012).

[8] Wikipedia contributors, "Crowdsourcing," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Crowdsourcing (accessed July 2012).

[9] Daren C. Brabham. "Crowdsourcing as a model for problem solving: Leveraging the collective intelligence of online communities for public good." PhD dissertation, University of Utah, 2010.

[10] Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. "Clasificación de iniciativas de crowdsourcing basada en tareas". *El profesional de la información*, 2012 (in press).

Lots of taxonomies, with little agreement about terminology, a reflection, I suspect, of the newness of crowdsourcing.

Typical cultural heritage crowdsourcing projects (OCR text correction, transcription, tagging) fall into the category of -- depending on whose typology one uses -- microwork (Wikipedia) / distributed human intelligence tasking (Brabham) / crowdproduction (Estellés).

3. Crowdsourcing projects

As we see above, two years ago the number of crowdsourcing projects was far fewer (~34) than today, and the number of cultural heritage crowdsourcing projects even fewer (2). Even though the number of cultural heritage projects is still not great, we must choose the ones to describe or risk writing a paper lengthier than anyone wants to read. Here we look in detail -- but not too much detail! -- at 3 crowdsourcing projects, two of which are newspapers project managed by libraries and one archival records project managed by a genealogy organization. We will mention a few others briefly.

3.1 Family Search[11] (http://www.familysearch.org)

Huge and prolific are adjectives quite appropriate for Family Search's project to index the world's historical records, including records from births, marriages, deaths, census, the military, land, and others from every continent but Antarctica. Since its launch in September 2005, Family Search volunteers have indexed more than 1,500,088,741 records and arbitrated more than 832,207,799 records. All records are freely searchable and available to anyone with Internet access.

As of July 2012, Family Search has over 780,000 registered users. In 2012 the average monthly number of active users and arbitrators is 90,215 and 9,174 respectively. The average monthly volume of records indexed is 48,935,686 and the average monthly records arbitrated is 23,891,167.

Other facts about Family Search:

- For the last 12 months (June 2011-May 2012) Family Search has had 278,906 active volunteers who have completed at least 1 record (this is higher than normal due to the 1940 census). Approximately 87,000 volunteers have done 40 records or less.

- For 2012 Family Search averaged 792,310 arbitrated records per day.

- Family Search publishes over 200 million indexed names per year (double entry with arbitration).

[11] Information and statistics are from private communications with Family Search.

- At present Family Search has more than 150 active projects. New projects are added weekly (current projects list is at http://indexing.familysearch.org/).

- Data entry and arbitration software user interface is available in English, Dutch, French, German, Italian, Japanese, Polish, Portuguese, Russian, Spanish, and Swedish. More languages and international projects coming.

Family Search indexing uses a custom-built Java-based workflow engine and Java webstart client software. Batches of record images are downloaded by volunteers to their own computers, transcribed, and, upon completion of the batch, uploaded to the Family Search indexing workflow engine.

Each record is transcribed by 2 volunteers. Arbitration is the process of reviewing and reconciling the differences between how the 2 volunteers interpreted the record. When the differences have been reconciled, the record is ready to publish.

3.2 Australian Newspapers Digitisation Program (http://trove.nla.gov.au/)

A paper about crowdsourced cultural heritage project would not be complete without mention of the National Library of Australia's newspaper digitisation program, the results of which can be found in its Trove (http://trove.nla.gov.au/). The National Library pioneered crowdsourced OCR text correction of historical newspapers.

As of June 2012, 68,908,757 lines of newspaper text had been corrected. Total registered users as of June 2012 63,553. During the 1st 6 months of 2012 an average of approximately 245,000 lines of newspaper text were corrected each month by ~3500 active registered users.

For those who want to learn more about Australia's crowdsourcing, look for anything written about crowdsourcing by Rose Holley, former manager for Trove. Ms. Holley now works at the National Archives of Australia.

3.3 California Digital Newspaper Collection (http://cdnc.ucr.edu)

The University of California Riverside has participated in the Library of Congress National Digital Newspaper Program since the program began in 2005. As of June 2012, CDNC has digitized 55,970 issues of newspaper comprising nearly 500,000 pages. On the CDNC website, newspapers are indexed and searched by article; at the Library of Congress's Chronicling America website, the same newspapers are indexed and searched by page.

In August 2011 CDNC added the user OCR text correction module to its Veridian digital library software. As of July 2012, 297 active registered users have corrected more than 395,000 lines of text.

3.4 Other projects

A report about newspaper crowdsourcing projects must mention Microtask and the National Library of Finland's *Digitalkoot* (http://www.digitalkoot.fi) for correcting OCR text to high accuracy. *Digitalkoot* is different from other projects mentioned here in that the useful work done is disguised as a game. As of July 2012 *Digitalkoot* gamers have contributed 407,734 minutes of their time to correcting OCR text. Since there are *Digitalkoot* project members at this meeting, I will leave any further discussion to them as they are far more knowledgeable about it than I am.

Distributed Proofreaders / Project Gutenberg (http://www.pgdp.net) is the oldest of crowdsourced cultural heritage projects and perhaps the oldest active crowdsourcing project today. It was started by Charles Franks in 2000 and became officially affiliated with Project Gutenberg in 2002. As of July 8, 2012 its volunteer proofreaders have added more than 40,000 public domain texts to Project Gutenberg. (cf. http://www.pgdp.net/).

The goal of the New York Public Library's *What's on the menu?* (http://menus.nypl.org/) is to transcribe dishes in approximately 45,000 menus dating from 1840 to the present. To date (July 2012) 974,329 dishes from 14,640 menus have been transcribed.

Transcribe Bentham (http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham) is a crowdsourcing project managed by the University College London. "Its aim is to engage the public in the online transcription of original and unstudied manuscript papers written by Jeremy Bentham (1748-1832), the great philosopher and reformer[12].

The US National Archives and Records Administration recently began a pilot project, the *Citizen Archivist Dashboard* (http://www.archives.gov/citizen-archivist/), to transcribe documents from the late 18th century to the 20th century. The documents include letters, presidential records, suffrage petitions, and fugitive slave case files and are classified by level of difficulty: beginner, intermediate, and advanced.

[12] Transcribe Bentham webpage http://www.ucl.ac.uk/transcribe-bentham/ University College London (accessed July 2012).

4. Crowdsourcing motivations

Digital historical newspaper collections are popular with genealogists. Several years ago the National Library of New Zealand surveyed users of its Papers Past collection and found that more than 50% used Papers Past for family history research[13]. Similarly in a 2010 report about Trove users, the National Library of Australia found that 50% of its users are family history researchers[14] and that more than 1/2 are 55 years of age or older.

A March-April 2012 survey showed that approximately 70% of the visitors to Utah Digital Newspapers used the collection for genealogical research[15]. At the meeting of the Newspaper Interest Group during the 2012 ALA Annual Conference, Brian Geiger[16] reported similar results from an informal survey of users of the text correction feature of the California Digital Newspapers Collection: Of the 136 users who answered the survey, more than 75% are 50 years old or older and nearly 1/2 used CDNC for genealogy or family history.

Family Search Indexing is obviously used mostly for genealogy and family history, and, although Family Search has not published a survey of its indexers and arbitrators similar to CDNC, New Zealand, and Australia, it is a safe bet that its users will also be mostly genealogists and family historians.

What do users themselves say about text correction? In Rose Holley's "Many hands make light work"[17] the National Library of Australia's Trove text correctors report

- "I enjoy the correction – it's a great way to learn more about past history and things of interest whilst doing a 'service to the community' by correcting text for the benefit of others."

- "We are sick of doing housework. We do it because it's addictive. It helps us and other people."

[13] Private communication with Tracy Powell, National Library of New Zealand.

[14] Libraries Australia Advisory Committee. "Trove report 2010." http://www.nla.gov.au/librariesaustralia/files/2011/11/laac-paper-2010-2-08-trove-report.pdf. (accessed July 2012).

[15] John Herbert and Randy Olsen. "Small town papers: Still delivering the news." Paper presented at the 2012 IFLA General Conference, Helsinki, Finland, August 11-17, 2012.

[16] Brian Geiger. "Improving the California Digital Newspaper Collection Software [SurveyMonkey survey]." Paper presented at the 2012 ALA Annual Conference, Anaheim, California USA, June 21-26, 2012.

[17] Rose Holley. "Many Hands Make Light Work." National Library of Australia (http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf) March 2009.

• "I have recently retired from IT and thought that I could be of some assistance to the project. It benefits me and other people. It helps with family research."

• "I enjoy typing, want to do something useful and find the content interesting."

Two text correctors for the California Digital Newspaper Collection say the following[18]

• "I am interested in all kinds of history. I have pursued genealogy as a hobby for many years. I correct text at CDNC because I see it as a constructive way to contribute to a worthwhile project. Because I am interested in history, I enjoy it."

• "I only correct the text on articles of local interest - nothing at state, national or international level, no advertisements, etc. The objective is to be able to help researchers to locate local people, places, organizations and events using the on-line search at CDNC. I correct local news & gossip, personal items, real estate transactions, superior court proceedings, county and local board of supervisors meetings, obituaries, birth notices, marriages, yachting news, etc."

In his book *Cognitive surplus: Creativity and generosity in a connected age*[19] Clay Shirky hypothesizes that people are now learning to use their free time for creative activities rather than consumptive ones as has been the trend since 1940. With some back-of-the-envelope calculations, Mr. Shirky estimates that the total human cognitive effort in creating all of Wikipedia in every language is about one hundred million hours. Furthermore he points out that Americans alone watch two hundred billion hours of TV every year, or enough time, if it would be devoted to projects similar to Wikipedia, to create about 2000 of them.

Although the comments from Trove and CDNC users are by no means scientific proof, it seems that these users are willing to devote some of their free time to a creative activity that benefits others: More accurate text in historical newspaper articles. Furthermore only 1 of Jeff Howe's 5 principles about the new labor pool can be applied to Trove and CDNC text corrections: The crowd (text correctors) is dispersed. Howe's 2nd principle -- the crowd has a short attention span -- applies to *some* of the text correctors, the ones who correct a few lines and don't ever re-visit but certainly does not apply to those who routinely correct 1000's of lines every month.

In addition to Shirky's book, there are a number of blogs and papers written about the motivations of crowdsourcing volunteers, intrinsic, extrinsic, for fun, to relieve boredom, for the good of the community, etc. Some are listed in the bibliography.

[18] Personal communication with CDNC text correctors.

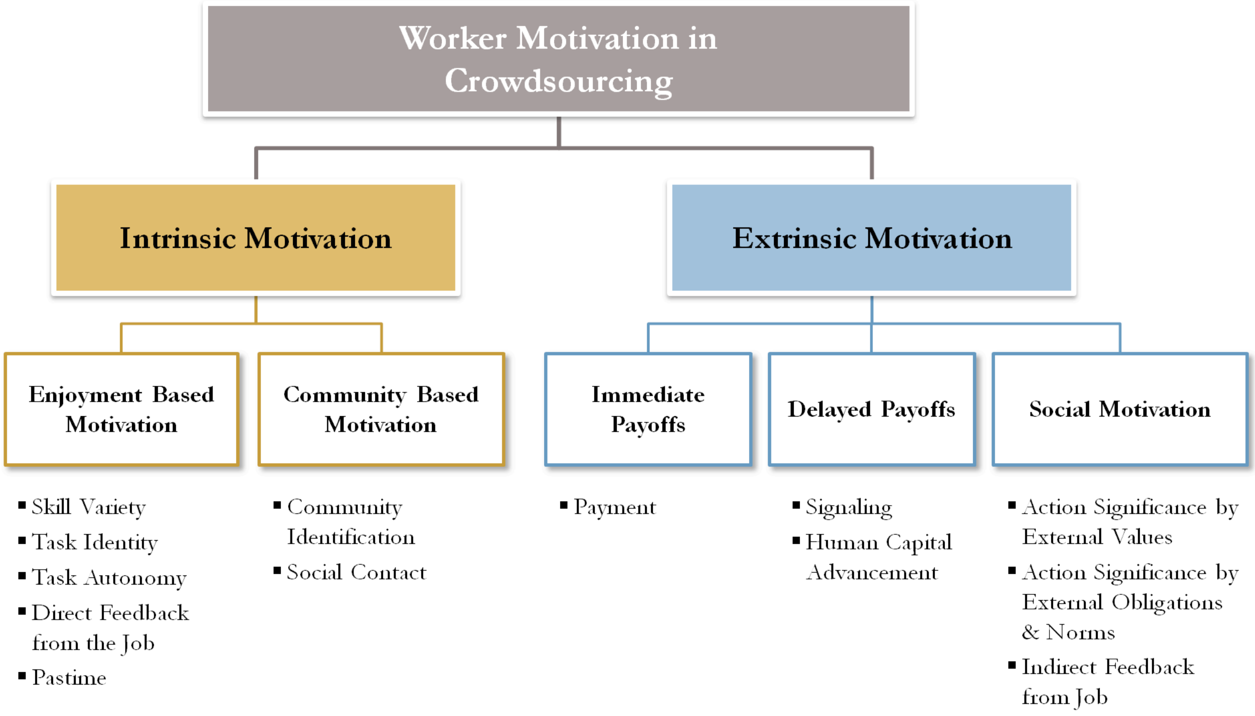
[19] Clay Shirky. *Cognitive surplus: Creativity and generosity in a connected age.* Penguin Press. New York. 2010.

For unpaid volunteers, perhaps most important motivations -- in my view -- is that crowdsourcing is simple and fun, gives a sense of community, provides a benefit to the broader community (not just crowdsourcing volunteers). In the August 23, 2007, the Editor in Chief's blog "Increase Motivation" at Pick the Brain[20] lists the following as motivational factors. Many of these motivational factors are obvious but a reminder is helpful. Trove's Rose Holley also cites this blog in her discussion of users' motivational factors in her paper *Many Hands Make Light Work*.

1. Consequences – Never use threats. They'll turn people against you. But making people aware of the negative consequences of not getting results (for everyone involved) can have a big impact. This one is also big for self motivation. If you don't get your act together, will you ever get what you want?
2. Pleasure – This is the old carrot on a stick technique. Providing pleasurable rewards creates eager and productive people.
3. Performance incentives – Appeal to people's selfish nature. Give them the opportunity to earn more for themselves by earning more for you.
4. Detailed instructions – If you want a specific result, give specific instructions. People work better when they know exactly what's expected.
5. Short and long term goals – Use both short and long term goals to guide the action process and create an overall philosophy.
6. Kindness – Get people on your side and they'll want to help you. Piss them off and they'll do everything they can to screw you over.
7. Deadlines – Many people are most productive right before a big deadline. They also have a hard time focusing until that deadline is looming overhead. Use this to your advantage by setting up a series of mini-deadlines building up to an end result.
8. Team Spirit – Create an environment of camaraderie. People work more effectively when they feel like part of team — they don't want to let others down.
10. Recognize achievement – Make a point to recognize achievements one-on-one and also in group settings. People like to see that their work isn't being ignored.
11. Personal stake – Think about the personal stake of others. What do they need? By understanding this you'll be able to keep people happy and productive.
12. Concentrate on outcomes – No one likes to work with someone standing over their shoulder. Focus on outcomes — make it clear what you want and cut people loose to get it done on their own.
13. Trust and Respect – Give people the trust and respect they deserve and they'll respond to requests much more favorably.
14. Create challenges – People are happy when they're progressing towards a goal. Give them the opportunity to face new and difficult problems and they'll be more enthusiastic.
15. Let people be creative – Don't expect everyone to do things your way. Allowing people to be creative creates a more optimistic environment and can lead to awesome new ideas.
16. Constructive criticism – Often people don't realize what they're doing wrong. Let

[20] Adapted from "Increase motivation." Pick the Brain (accessed July 2012 http://www.pickthebrain.com/blog/21-proven-motivation-tactics).

them know. Most people want to improve and will make an effort once they know how to do it.
17. Demand improvement – Don’t let people stagnate. Each time someone advances raise the bar a little higher (especially for yourself).
18. Make it fun – Work is most enjoyable when it doesn’t feel like work at all. Let people have fun and the positive environment will lead to better results.
19. Create opportunities – Give people the opportunity to advance. Let them know that hard work will pay off.
20. Communication – Keep the communication channels open. By being aware of potential problems you can fix them before a serious dispute arises.
21. Make it stimulating – Mix it up. Don’t ask people to do the same boring tasks all the time. A stimulating environment creates enthusiasm and the opportunity for “big picture” thinking.


Worker Motivation in Crowdsourcing
Intrinsic Motivation
Extrinsic Motivation
Enjoyment Based Motivation
Community Based Motivation
Immediate Payoffs
Delayed Payoffs
Social Motivation
▪ Skill Variety
▪ Task Identity
▪ Task Autonomy
▪ Direct Feedback from the Job
▪ Pastime
▪ Community Identification
▪ Social Contact
▪ Payment
▪ Signaling
▪ Human Capital Advancement
▪ Action Significance by External Values
▪ Action Significance by External Obligations & Norms
▪ Indirect Feedback from Job

Peter Organisciak, PhD student at the University of Illinois School of Library and Information Science who studies and writes about crowdsourcing, lists similar crowd motivators in his blog post “Motivation of crowds: The incentives that make crowdsourcing work”[21]: (1) money, (2) fun, (3) boredom, (4) achievement, (5) charity, (6) academia, (7) participation, (8) self-benefit, (9) forced, and (10) interest.

[21] Peter Organisciak. Crowdstorming blog. “Motivation of corwds: The incentives that make crowdsourcing work.” January 31, 2008. (accessed at http://crowdstorming.wordpress.com/2008/01/31/motivation-of-crowds-the-incentives-that-make-crowdsourcing-work/)

A final and theoretical word about crowdsourcing motivation is given by Kaufmann *et al*[22] and summarized in the graphic above. Motivation theory divides human motivation into *intrinsic* and *extrinsic*. Intrinsic motivation "refers to motivation that is driven by an interest or enjoyment in the task itself, and exists within the individual rather than relying on any external pressure."[23] On the other hand extrinsic motivation "refers to performance of an activity in order to attain an outcome."[24] If the Trove and CDNC volunteers' reports above are any indication, intrinsic motivation is dominant motivator for cultural heritage crowdsourcing projects.

5. Crowdsourcing benefits

Crowdsourcing, like most things, has both value and cost. Some aspects of crowdsourcing are easy to measure or quantify, for example, counting the number of characters of lines corrected, the number of registered / active users, duration of visits to the website, costs, and the like. Other aspects, especially those of less tangible value, are more difficult. Let's look at the easy stuff.

5.1 Avoided costs

The obvious benefit from crowdsourced OCR text correction or transcription is improved search. This is especially important for digitized newspaper collections because OCR text accuracy is often very poor. Edwin Kiljin reports raw OCR character accuracies of 68% for early 20th century newspapers[25]. For a sample of 45 pages of Trove digitized newspapers from 1803 to 1954, Rose Holley reports that raw OCR character accuracy varied from 71% to 98%[26].

Of course correction of OCR text can also be outsourced to service bureaus. Australia, New Zealand, Singapore, CDNC, and others do rely on outsourced OCR text correction, but,

[22] Nicholas Kaufmann, Thimo Schulze, and Daniel Veit. "More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk." Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4th-7th 2011 (accessed July 2012 at http://schader.bwl.uni-mannheim.de/fileadmin/files/publikationen/Kaufmann_Schulze_Veit_2011_-_More_than_fun_and_money_Worker_motivation_in_Crowdsourcing_-_A_Study_on_Mechanical_Turk_AMCIS_2011.pdf)

[23] Wikipedia contributors, "Motivation," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Motivation (accessed July 2012).

[24] Ibid.

[25] Edwin Kiljin. "The current state-of-art in newspaper digitization." D-Lib Magazine. January/February 2008. (Accessed at http://www.dlib.org/dlib/january08/klijn/01klijn.html).

[26] Rose Holley. "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine. March/April 2009. (Accessed at http://www.dlib.org/dlib/march09/holley/03holley.html).

because it is costly, correction is limited to article headlines or, in the case of Trove, to headlines plus the 1st 4 lines of certain articles.

Let's do a back-of-the-envelope calculation using CDNC and Trove's count of corrected lines of newspaper text. Depending on the era of the newspaper, the number of columns, font size, and layout, there are 25 to 50 characters per newspaper column line. Let's assume 35 characters per line.

Depending on labor costs at the service bureau, outsourced text correction to 99.5% accuracy costs range from USD $0.35 per 1000 characters to more than USD $1.00 per 1000 characters. For this back-of-the-envelope calculation, et's assume USD $0.50 per 1000 characters.

As of July 12, 2012, volunteers at CDNC have corrected 394.365 lines of text. Using these assumptions, the value of CDNC volunteer labor is 394,365 lines x 35 characters x 1/1000 characters x $0.50 = $6901. A similar calculation for the National Library of Australia's Trove, where volunteers have corrected 69,918,892 lines of text (July 12, 2012), results in a value of Trove volunteer labor of $1,223,581.

These numbers are truly significant[27]! However as we shall see below, the monetary value of volunteer labor may not be crowdsourcing's most significant benefit.

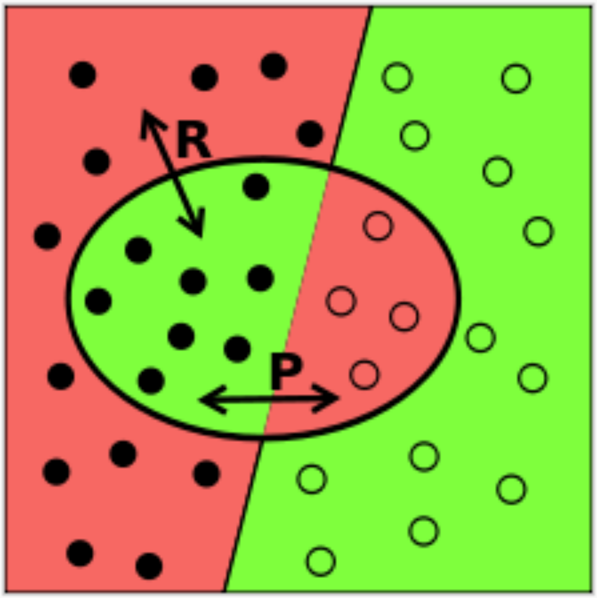
5.2 Unavoidable costs

What does implementing a crowdsourcing application cost? The obvious costs are made up of things like the (1) crowdsourcing software itself, (2) software and hardware infrastructure, (3) administration and support of the crowdsourcing software and the software and hardware infrastructure, (4) Internet and communications. Less obvious but no less important are (1) crowdsourcing volunteer support (help desk), (2) documentation and FAQs on how to use the crowdsourcing application, (3) marketing or letting the world know about the crowdsourcing application, and (4) best practices documentation and guidelines.

These costs are very dependent on prevailing wages and the size of the crowdsourcing project. For example, Family Search is a large division of an even larger organization with a budget of millions of dollars. On the other hand CDNC's Veridian OCR text correction and digital library software is managed with less than 20% of a system administrator's time and with an annual budget of less than USD $20,000.

Crowdsourcing software costs are easy to calculate for commercial off-the-shelf software (COTS) but difficult to calculate for custom-built software. And most cultural heritage crowdsourcing software appears to be custom-built at present.

[27] If you don't like the assumptions, put numbers you like into the following formula: *linesCorrected* x *charactersPerLine*/1000 x *costPer1000Characters*.



In this figure the relevant items are to the left of the straight line while the retrieved items are within the oval. The red regions represent errors. On the left these are the relevant items not retrieved (false negatives), while on the right they are the retrieved items that are not relevant (false positives). **Precision** and **recall** are the quotient of the left green region by respectively the oval (horizontal arrow) and the left region (diagonal arrow).

A very crude back-of-the-envelope estimation of costs: An unlimited license for Veridian digital library plus OCR text correction software is approximately USD $33,000 per year with software costs amortized over 3 years[28]. Estimating the obvious and less obvious costs (included from obvious costs items 2, 3, and 4, included from less obvious costs items 1, 2 3, and 4) from above as 50% of this amount, this gives an estimated annual cost of $49,500. If this cost is compared to the avoided costs of text correction for Trove from above, crowdsourced OCR text correction is a bargain.

Remember this is a very rough estimate! Actual costs may be significantly less for small projects with limited licenses or significantly more for custom-built software for large projects. In particular maintenance costs for custom-built software will with near certainty be higher.

5.3 Improved accuracy

What does more accurate OCR text mean for search? One must realize that the accuracy of raw OCR text varies widely and is often quite poor (see remarks by Edwin Kiljin and Rose Holley reported in section 5.1). For the purpose of this discussion, let's (optimistically) assume an average raw OCR character accuracy of 90%.

The average length of a word in the English language is 5 characters. This means that words have an average accuracy of 90% x 90% x 90% x 90% x 90% = 59% or that only 6 words out of 10 in raw OCR text are correct. Even optimistically assuming average raw OCR character accuracy is 95% still gives an average word accuracy of only 77%.

And since the average length includes stop words like *a*, *the*, *and*, etc, the average length of "interesting" words for search -- for example, personal, place, and organization names -- will be longer and their accuracy even lower.

[28] Includes cost of Veridian digital library software and OCR text correction module amortized over 3 years. Optional support fees are about 15% of cost and included in the estimated cost.

In information retrieval, *precision* is the fraction of retrieved objects that are relevant to the search and *recall* is the fraction of relevant objects that are retrieved (see figure[29]). A perfect score for *precision* and *recall* is 1.0. Perfect *precision* (1.0) means that nothing irrelevant is retrieved; perfect *recall* (1.0) means that everything relevant is retrieved. Searches with low *precision* are a nuisance if one must sort through many irrelevant documents, but searches with low *recall* are an anathema to genealogists. What does this mean? For example, if my grandmother's family name 'Arndt' occurs on 10 pages at Chronicling America, but, if we assume 90% raw OCR text accuracy, a search will find only 6 pages and recall is 6/10 = 0.6[30].

To my knowledge no one has yet measured the accuracy of crowdsourced corrected OCR text (or crowdsourced transcribed handwritten manuscripts). However, operators at service bureaus routinely correct OCR text to 99.5% so it's reasonable to assume that the same accuracy is possible for crowdsourced OCR text correction. Furthermore genealogists, who apparently comprise the majority of volunteer text correctors, will be particularly careful to correct names. Character accuracy of 99.5% raises the accuracy of an average length English language word to 97.5% and would raise the *recall* of the Chronicling America search for 'Arndt' to nearly 1.

5.4 <u>The real benefit of crowdsourcing</u>

In his blog Trevor Owens forcefully argues that the real benefit of crowdsourcing OCR correction or transcription is not improved search or searches with higher *precision* and *recall*. Instead he believes that the real benefit is the meaningful activity and the facility for purposeful contributions that it provides for volunteers. (This is one of those difficult to measure and quantified values mentioned above.) Here's an excerpt from his excellent blog on the objectives of crowdsourcing[31]:

> <u>Crowdsourcing is better at Digital Collections than Displaying Digital Collections</u>
>
> What crowdsourcing does, that most digital collection platforms fail to do, is offers 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. Far from being an instrument which enables us to ultimately better deliver content to end

[29] Wikipedia contributors, "Precision and recall," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Precision_and_recall (accessed July 2012).

[30] My grandmother's maiden name is actually found 9,534 times on 477 pages at Chronicling America (July 2012). I have no idea if all occurrences are accurate and have even less idea about the total number of times 'Arndt' is found in Chronicling America. I used 6 and 10 because it makes the math easy.

[31] Trevor Owens. "Crowdsourcing cultural heritage: The objectives are upside down." Blog posted March 10, 2012 at http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/

> users, crowdsourcing is the best way to actually engage our users in the fundamental reason that these digital collections exist in the first place.
>
> Meaningful Activity is the Apex of User Experience for Cultural Heritage Collections
>
> When we adopt this mindset, the money spent on crowdsourcing projects in terms of designing and building systems, in terms of staff time to manage, etc. is not something that can be compared to the costs of having someone transcribe documents on mechanical turk. Think about it this way, the transcription of those documents is actually a precious resource, a precious bit of activity that would mean the world to someone.

How does one value the opportunity to contribute to the public memory through crowdsourcing OCR text correction, manuscript transcription, or genealogical records data entry? As I pointed out above, this is one of the intangible, difficult-to-quantify benefits. However although intangible and difficult to measure, it should by no means be ignored or treated as insignificant. Mr. Owens arguments add to the other value of crowdsourcing cultural heritage collections -- improved search. Even though difficult to measure or quantify, crowdsourcing in a very simple way increases libraries' relevance to the communities they serve in the age of the Internet.

6. Conclusion

When I began to write this paper, I had not expected to find so much interest in crowdsourcing at cultural heritage organizations. But as the numbers of crowdsourcing projects listed by Wikipedia in 2010 and today shows, interest in crowdsourcing has proliferated, both in numbers of projects and in realms of its application. For example, crowdsourcing is used by Netflix to improve its video recommendations algorithms, by a Canadian mining company to find gold, by Kickstarter to raise money for innovative projects (including library projects!), and by Galaxy Zoo to classify a million galaxies from the Sloan Digital Sky Survey. Should one not expect that cultural heritage organizations can use crowdsourcing to enhance their value to their local communities and to the world's cultural heritage?

A few cultural heritage crowdsourcing projects were briefly described earlier in this paper. How many others projects are planned by libraries, archives, and museums? How many projects are already active but undiscovered, at least undiscovered by me?

Web 2.0 and crowdsourcing give libraries some truly interesting opportunities, opportunities which can engage their patrons as never before.

Bibliography

1. Libraries Australia Advisory Committee. "Trove report 2010." (Accessed July 2012 at http://www.nla.gov.au/librariesaustralia/files/2011/11/laac-paper-2010-2-08-trove-report.pdf.).

2. Daren C. Brabham. “Crowdsourcing as a model for problem solving: Leveraging the collective intelligence of online communities for public good.” PhD dissertation, University of Utah, 2010.

3. Crowdsourcing your library program challenges. Wiki at http://crowdsourcing-librarychallenges.wikispaces.com/

4. Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. *Towards an integrated crowdsourcing definition.* Journal of Information Science XX(X). 2012.

5. Brian Geiger. “Improving the California Digital Newspaper Collection Software [SurveyMonkey survey].” Paper presented at the 2012 ALA Annual Conference, Anaheim, California USA, June 21-26, 2012.

6. John Herbert and Randy Olsen. “Small town papers: Still delivering the news.” Paper presented at the 2012 IFLA General Conference, Helsinki, Finland, August 11-17, 2012.

7. Rose Holley. Rose Holley’s Blog - Views and news on digital libraries and archives. http://rose-holley.blogspot.com/

8. Rose Holley. “How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazin. March/April 2009. (Accessed at http://www.dlib.org/dlib/march09/holley/03holley.html

9.Rose Holley. “Many Hands Make Light Work.” National Library of Australia (http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf) March 2009.

10. Jeff Howe. “The rise of crowdsourcing.” *Wired,* Issue 14.06, June 2006.

11. Nicholas Kaufmann, Thimo Schulze, and Daniel Veit. “More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk.” Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4th-7th 2011 (Accessed July 2012 at http://schader.bwl.uni-mannheim.de/fileadmin/files/publikationen/Kaufmann_Schulze_Veit_2011_-_More_than_fun_and_money_Worker_motivation_in_Crowdsourcing_-_A_Study_on_Mechanical_Turk_AMCIS_2011.pdf).

12. Edwin Kiljin. “The current state-of-art in newspaper digitization.” D-Lib Magazine. January/February 2008. (Accessed at http://www.dlib.org/dlib/january08/klijn/01klijn.html).

13. Christine Madsen Blog. http://christinemadsen.com/.

14. Matt Mickiewicz. Crowdsource your success. Presentation at Affiliate Summit East 2012. (Accessed July 2012 at http://www.slideshare.net/affsum/crowdsource-your-success).

15. Peter Organisciak. Crowdstorming blog. “Motivation of corwds: The incentives that make crowdsourcing work.” January 31, 2008. (Accessed July 2012 at http://crowdstorming.wordpress.com/2008/01/31/motivation-of-crowds-the-incentives-that-make-crowdsourcing-work/),

16. Trevor Owens Blog. User Centered Digital History. http://www.trevorowens.org/

17. Trevor Owens. “Crowdsourcing cultural heritage: The objectives are upside down.” Blog posted March 10, 2012 at http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/.

18. Pick the Brain blog. (Accessed July 2012 http://www.pickthebrain.com/blog/21-proven-motivation-tactics).

19. Nicole Saylor. Crowdsourcing library collections. Presentation at ARL Fall Forum, October 14, 2011. (Accessed July 2012 at http://www.arl.org/bm~doc/ff11-saylor.pdf).

20. Clay Shirky. *Cognitive surplus: Creativity and generosity in a connected age*. Penguin Press. New York. 2010.

21. James Surowiecki. *The wisdom of crowds*. New York: Random House. 2004.

22. Wikipedia contributors, "Crowdsourcing," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Crowdsourcing. (Accessed June 2012).

23. Wikipedia contributors, "List of crowdsourcing projects," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects (Accessed July 2012).

24. Wikipedia contributors, "Motivation," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/wiki/Motivation (Accessed July 2012).