IFLA WLIC 2013
SINGAPORE

Crowdsourcing the world's cultural heritage: Part II

Frederick Zarndt
IFLA Newspapers Section, Coronado CA USA
frederick@frederickzarndt.com

Brian Geiger
California Digital Newspapers Collection, University of California Riverside, Riverside CA USA
bgeiger@ucr.edu

Alyssa Pacy
Cambridge Public Library, Cambridge MA USA
apacy@cambridgema.gov

Stefan Boddie
DL Consulting, Hamilton, New Zealand
stefan@dlconsulting.com

Abstract

Wikipedia, founded in 2001, contains more than 20,000,000 articles in 282 languages written and edited by 100,000 volunteers from the worldwide "crowd". Open source software such as Apache, MySQL, Linux, Postgresql, Mozilla FireFox, and many, many others "proved that a network of passionate, geeky volunteers could write code just as well as the highly paid developers at Microsoft or Sun Microsystems." (Wired Magazine, June 2006). Since its inception in April 2009, Kickstarter raised over $350,000,000 for 30,000 projects from more than 2,500,000 members of the global "crowd".

In all of its many flavors, crowdsourcing works. It works for cultural heritage organizations too. At the 2012 Mikkeli Finland IFLA satellite conference, two of the authors described results of crowdsourced OCR text correction for the National Library of Australia's Trove and the California Digital Newspaper Collection (CDNC) as well as the astounding number of historical birth, death, marriage, census, and other records transcribed by "crowd" volunteers at Family Search.

In this paper we add the digital newspaper collection at the Public Library in Cambridge Massachusetts (population 105,000) to the cultural heritage digital historical newspaper collections mix and examine in more detail

1. *Demographics: What is the age of the "crowd", their profession, gender, and how did they learn about text correction?*
2. *Experiences: What do text correctors like about newspapers and about the text correction? Excerpts from user interviews will be presented.*
3. *Motivation: What makes users correct text? As with "experiences", the authors will excerpt user interviews and present academic crowdsourcing motivational research results.*
4. *Quality: OCR of newspaper text is often of very poor quality. What is the quality of the crowdsourced corrected text? What effect does increased text accuracy have on search recall? Before and after measurements of text accuracy will be presented.*
5. *Preferred data: What types of newspaper stories do text correctors prefer?*
6. *Economics: What is the estimated economic value of corrected text? What are the costs of providing a text correction infrastructure? Real financial measures from Trove used in the National Library of Australia's annual report are presented.*
7. *Marketing: What are effective strategies for promoting crowdsourcing at libraries?*
8. *You will see that crowd sourcing is not only feasible but also practical and desirable. You will wonder why your own cultural heritage organization hasn't begun its own crowdsourcing project!*

Keywords: newspapers, digital historical newspaper collections, crowdsourcing, OCR text, text correction

---

1. Crowdsourcing and libraries

In recent years crowdsourcing has exploded. The word itself is much in vogue. On July 28, 2013, Wikipedia listed 127 references to other Wikipedia pages in its "Crowdsourcing" category whereas on January 22, 2010[1], the same page had only 41 references. Similarly on July 28, 2013, the Wikipedia page on "Crowdsourcing" itself lists 79 external references; on July 5, 2010, only 10 external references.

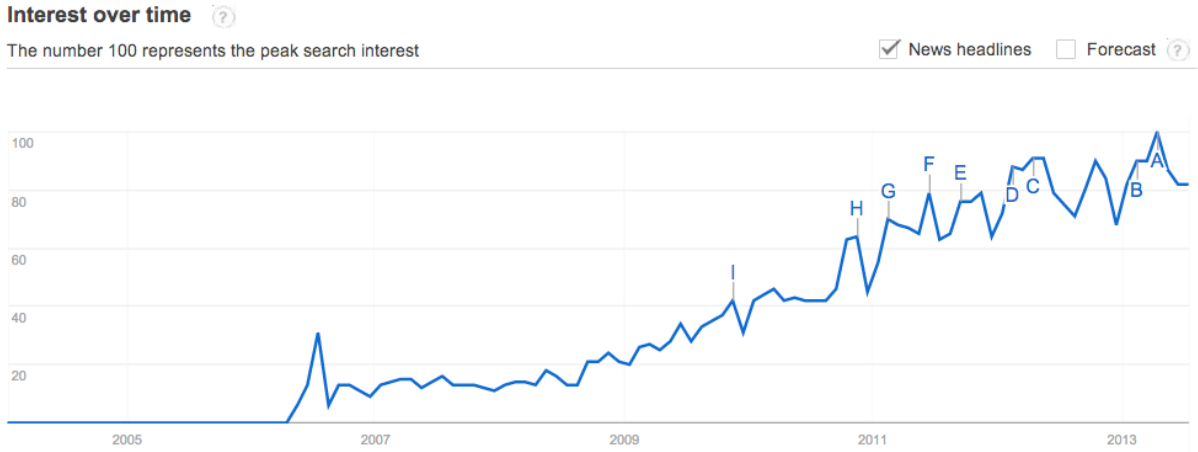
The word "crowdsourcing" was coined by Jeff Howe in the article *"The rise of crowdsourcing"* written for Wired magazine[2] in June 2006. In it Howe drafted 5 principles describing the new labor pool:

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).



Google Trends "crowdsourcing" 2004 to July 2013

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 Daren Brabham, who seems to be the first to define it in scientific literature[4]

> *crowdsourcing is an online, distributed problem-solving and production model.*

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*

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

[4] Daren Brabham. "*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. 2008. (http://www.clickadvisor.com/downloads/Brabham_Crowdsourcing_Problem_Solving.pdf accessed July 2013).

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

> *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.*

On January 25, 2010, Wikipedia listed 34 crowdsourcing projects[6]. In July 2013, that list had grown to ~168 projects[7]. Crowdsourced projects range from Wikipedia, the Open Dinosaur Project (http://opendino.wordpress.com/), a failed effort to purchase the Pabst Brewing Company, and many, many others, some of which practically beggar belief. And of the 168 projects listed, 11 are connected with digitized books, journals, manuscripts, or records, and libraries.

2. Demographics

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. In a similar 2010 report about Trove users, the National Library of Australia found that 50% of its users are family history researchers and that more than 1/2 are 55 years of age or older. And a March-April 2012 survey done by Utah Digital Newspapers showed that approximately 70% of the visitors to collection used it for genealogical research[8].

In order to learn about their user demographic the California Digital Newspaper Collection (CDNC) and the Cambridge Public Library surveyed users of their collections[9]. CDNC had surveyed its users in 2012, and, since that survey, CDNC has gained many new users. It wanted to see if demographic trends had shifted and to broaden the scope of the survey with additional questions. From February to May 2013 the CDNC user survey got 555 responses. From January to May 2013 the Cambridge survey got 30 responses[10]. For both collections users are overwhelmingly genealogists / family historians: 82% for Cambridge and 66% for CDNC.

---

[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.

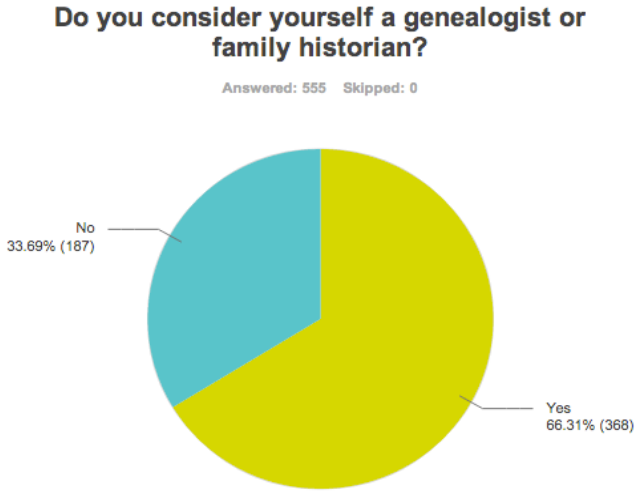
[7] Wikipedia contributors, "List of crowdsourcing projects," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects (accessed July 28, 2013).

[8] Randy Olsen and John Herbert. *Small town papers: still delivering the news*. World Library and Information Congress. Helsinki, Finland. August 2012. http://conference.ifla.org/past/ifla78/session-119 (accessed June 1, 2013).

[9] The CDNC survey questions can be found at the end of this paper. The Cambridge questions are nearly identical except for the name of the collection.

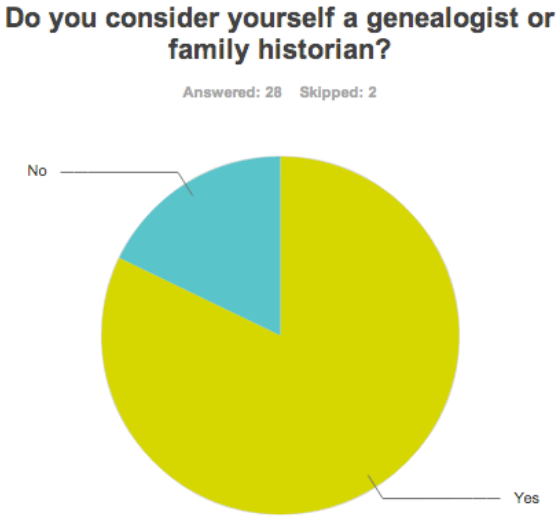
[10] The surveys are still online. As of July 20, 2013, the CDNC survey now has 643 respondents and the Cambridge survey 32.

As for age, 75% of Cambridge users are 50+ years old; nearly 80% of CDNC users are 50+ years old. Survey results also show that Cambridge had no survey respondents less than 30 years of age while fewer than 5% of CDNC respondents are under 30. The majority of Papers Past and Trove newspaper collection users are also 50+ years of age according to New Zealand and Australia library surveys.



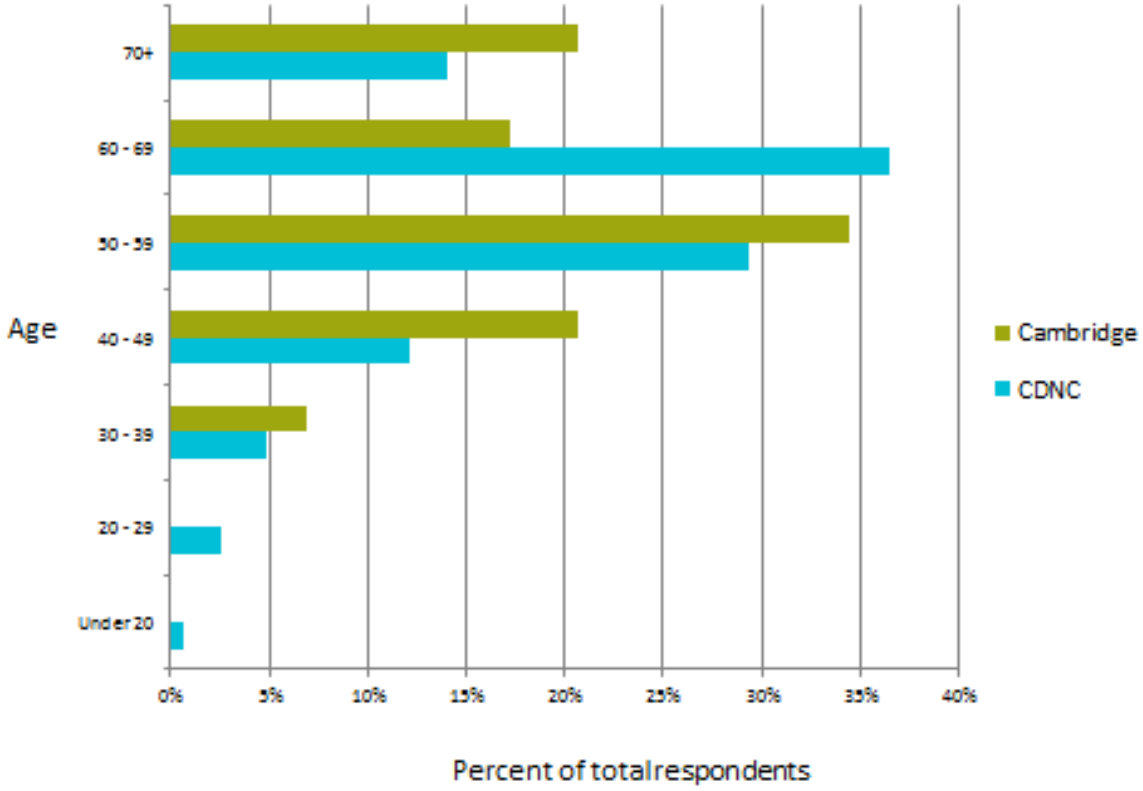
Answer Choices	Responses	
Yes	**66.31%**	368
No	**33.69%**	187
Total		555

CDNC User Demographic



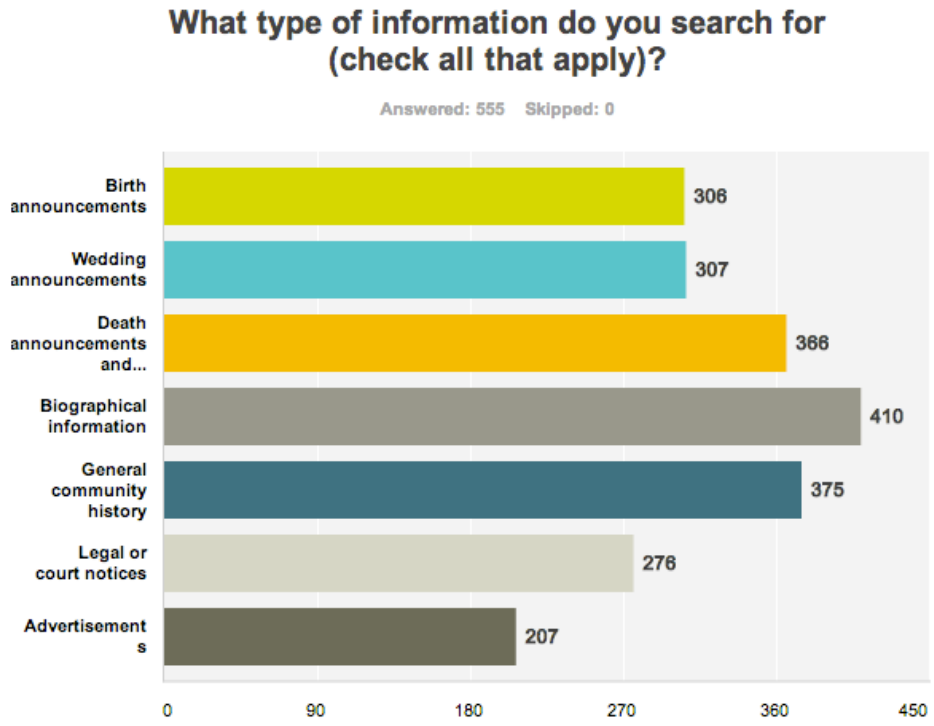
Answer Choices	Responses	
Yes	**82.14%**	23
No	**17.86%**	5
Total		28

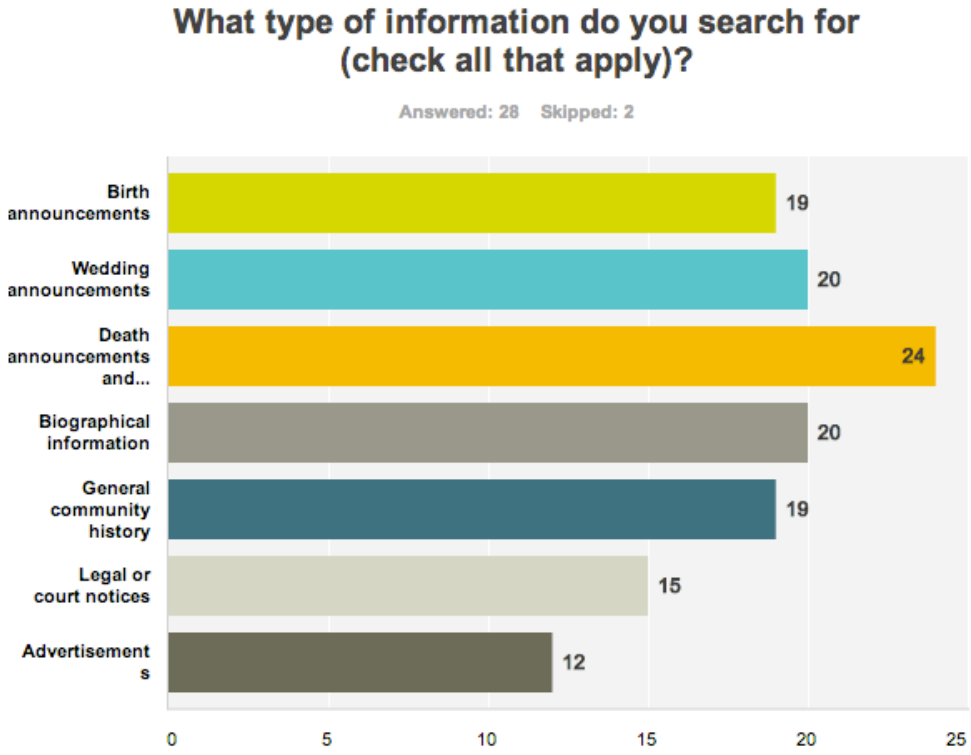
Cambridge Public Library User Demographic



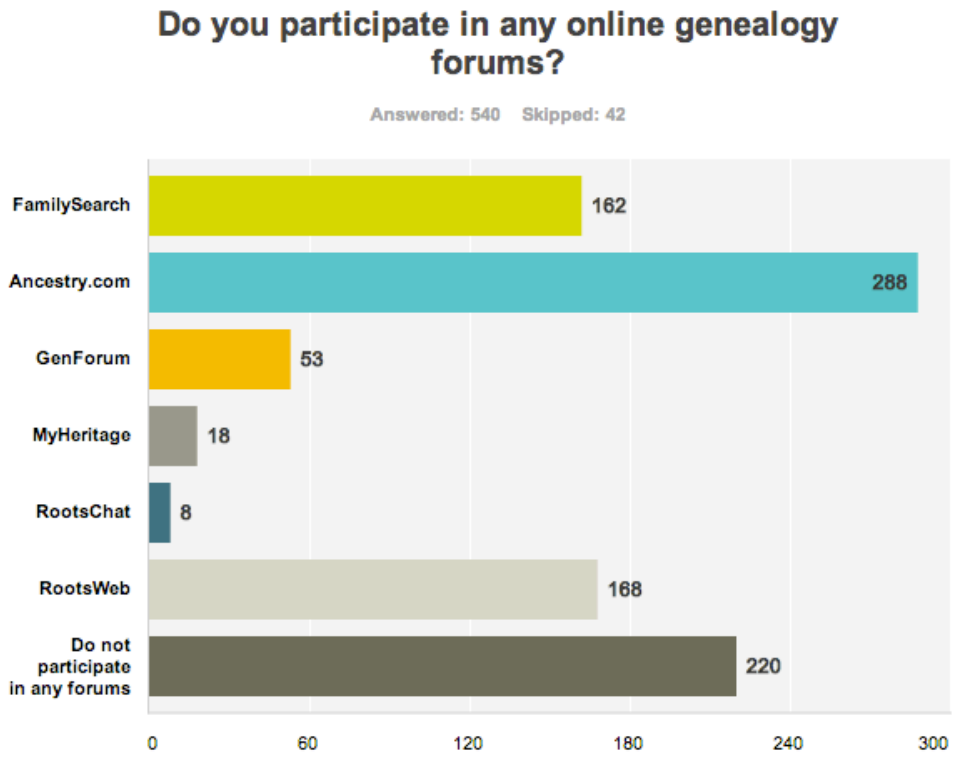
Both the CDNC and Cambridge surveys ask which material is most interesting to its users. The graphs below show what one would expect for users interested in genealogy: They search mostly for obituaries, general family announcements (births, weddings), and biographical information. Olsen and Herbert's Utah Digital Newspapers user surveys report similar results[11].

[11] Olsen and Herbert. *Small town papers: still delivering the news*.





The CDNC survey asks its users if they participate in or subscribe to a genealogy services like FamilySearch, Ancestry.com, RootsWeb, etc. Survey results show that about 60% use one or more one or more such services.



3. Motivations

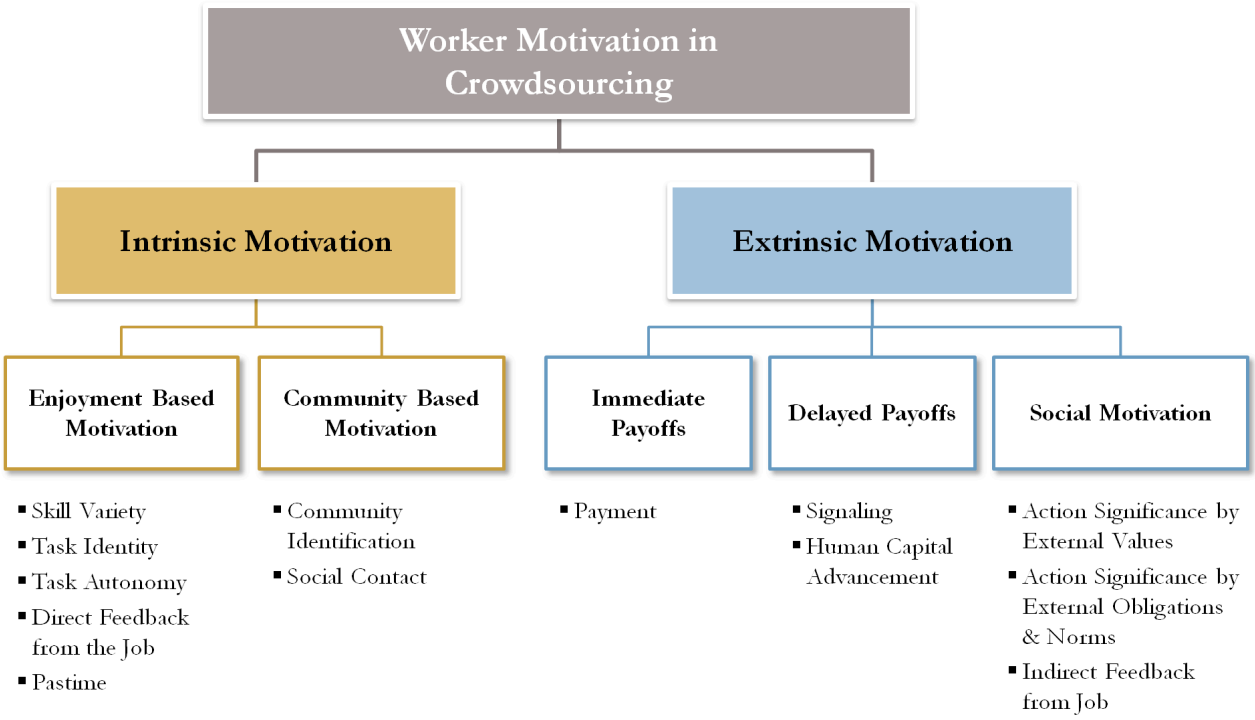
In his book *Cognitive surplus: Creativity and generosity in a connected age*[12] Clay Shirky hypothesizes that people are now learning to use their free time for creative activities rather than consumptive ones as has been the trend since 1940. With plausible, back-of-the-envelope calculations, Mr. 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.

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.

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"[13]: (1) money, (2) fun, (3) boredom, (4) achievement, (5) charity, (6) academia, (7) participation, (8) self-benefit, (9) forced, and (10) interest.

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

[13] 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 theoretical word about crowdsourcing motivation is given by Kaufmann et al[14] 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."[15] On the other hand extrinsic motivation "refers to performance of an activity in order to attain an outcome."[16] If the Trove, CDNC, and Cambridge volunteer reports below are any indication, intrinsic motivation is certainly the dominant motivator for cultural heritage crowdsourcing projects.

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."*

[14] Kaufmann, Nicolas, Thimo Schulze, and Daniel Veit, "More than fun and money: Worker Motivation in Crowdsourcing – A Study on Mechanical Turk". *AMCIS 2011 Proceedings*. http://aisel.aisnet.org/amcis2011_submissions/340.

[15] Wikipedia contributors. "Motivation". Wikipedia, The Free Encyclopedia. (http://en.wikipedia.org/wiki/Motivation accessed July 2013).

[16] Ibid.

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

"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."

And CDNC text correctors say this about correcting text:

"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."

Wesley, California

"I have always been interested in history, especially the development of the American West, and nothing brings it alive better than newspapers of the time. I believe them to be an invaluable source of knowledge for us and future generations."

David, United Kingdom

"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."

Ann, California

"CDNC is an excellent source of information matching my personal interest in such topics as sea history, development of shipbuilding, clippers and other ships etc. ... Unfortunately, the quality of text ... is rather poor I'm afraid. This is why I started to do all corrections necessary for myself ... and to leave the corrected text for use of others. .... I am not doing this very regularly as this is just my hobby and pleasure."

Jerzey, Poland

"I am correcting text for the Coronado Tent City Program for 1903. It is important to correct any problems with personal names and other information so that researchers will be able to search by keyword and be assured of retrieving desired results. ... type fonts cause a great deal of difficulty in digitizing the text and can cause problems for searchers. Also, many of the guests' names at Tent City and Hotel Del Coronado were taken from the registration books and reported in the Program. This led to many problems in spelling of last names and the editors were not careful to be consistent in the spellings. This Program is an important resource since it provides an excellent picture of daily life in Tent City and captures much of the history of Coronado itself."

Gene, California

And not to leave out Cambridge Public Library text correctors:

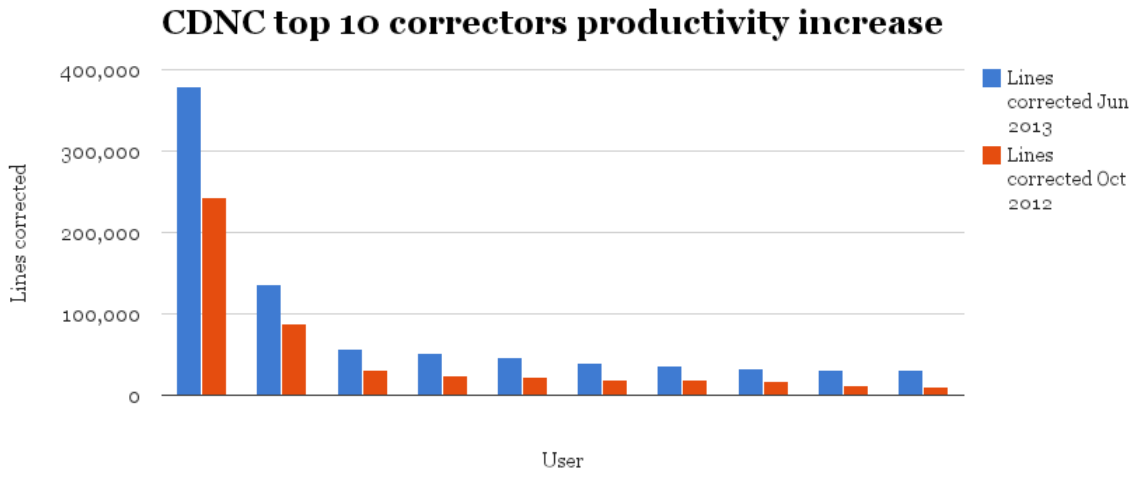
> *"As an amateur historical researcher my time for research is very limited. Making time to travel to archives, libraries, and historical societies does not happen as often as I would like. The Cambridge Public Library's online newspaper collection has been an invaluable resource and it is fun. I am very grateful for all the help I have received over the years from so many research organizations. Correcting text has several benefits. It makes it much more likely that I will find a story if I decide to search for it in the future. It is a way of saying 'thank you' to the Cambridge Library for having such a great resource available and maybe I can make the next person's research a little easier. It is my own little historical preservation project."*
>
> *Daniel, Somerville, Massachusetts, USA*

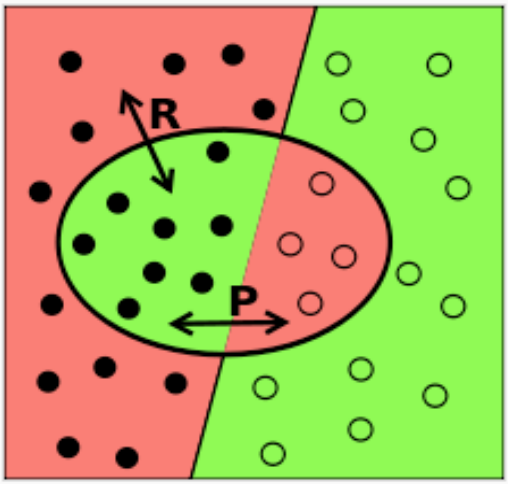
> *"Many of my paternal relatives are from Cambridge. While I had the basics (name, date of birth, addresses, etc.), I really didn't know that much about them. Reading the many newspaper articles has given me great insight into their daily and personal lives. For instance, I didn't know that my paternal great-grandfather, John Hargrave Kelsey, narrowly missed being a casualty of the San Francisco earthquake. I learned that both of my paternal grandfathers were minstrel show singers and dancers for their fraternal organizations – what a delightful surprise. My other paternal great-grandfather, William L. Boyson, was a noted and valued employee of Riverside Press for over 50 years. Previous research only showed me that he was a bookbinder. It was fun to read about a dog-bite law suit against a great-grand uncle. Many of the weddings of my relatives were reported in the papers, and it is fun to imagine just what extravagant events they must have been. Reading other articles about non-relatives and citizens and advertisements gives great insight into the times."*
>
> *Maude Marie, Cooper City, Florida, USA*

CDNC		Cambridge	
User	**Lines corrected**	**Lines corrected**	**User**
1	415,058	15,813	1
2	150,491	13,550	2
3	60,833	5,139	3
4	59,777	2,252	4
5	57,306	1,701	5
6	39,617	1,138	6
7	37,652	922	7
8	34,775	909	8
9	31,319	819	9
10	30,408	498	10

Although the comments from the text correctors 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, namely, 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, CDNC, and Cambridge 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.

How motivated are text correctors? The table above shows the number of lines of text corrected by the top 10 text correctors at CDNC and Cambridge (measured Jul 2013). As you can see, some text correctors are astoundingly productive! And as the bar chart of CDNC text correctors below shows, their productivity does not diminish with time, at least not yet.





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).

4. Benefits

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

4.1 Improved text accuracy

The most obvious benefit from crowdsourced OCR text correction or transcription is improved search. This is especially important for digitized newspaper collections because raw, uncorrected OCR text accuracy is often very poor. Edwin Kiljin reports raw OCR character accuracies of 68% for early 20th century newspapers[18]. 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%[19].

[18] 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).

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 above). 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 in raw OCR text of average character accuracy have an average word 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.

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 above[20]). 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 one of the author's grandmother's family name 'Arndt' occurs on 10 pages at Chronicling America, but, if we assume 90% raw OCR character accuracy, a search will find only 6 pages and recall is 6/10 = 0.6.

Let's look at the accuracy of raw OCR from several CDNC newspaper titles. Raw OCR errors from 176 lines in issues from various dates and pages were counted. To calculate word accuracy, we assume average English language word length (5 characters) and multiply raw OCR accuracy by itself 5 times.

Raw OCR character and word accuracy

Title	OCR character accuracy	~OCR word accuracy
PRP Pacific Rural Press 1871 - 1922	92.6%	68.1%
SFC San Francisco Call 1890 - 1913	92.6%	68.1%
LAH Los Angeles Herald 1873 - 1910	88.7%	54.9%
LH Livermore Herald 1877 - 1899	88.6%	54.6%
DAC Daily Alta California 1841 - 1891	88.2%	53.4%
CFJ California Farmer and Journal of Useful Sciences 1855 - 1880	86.5%	48.4%

[19] 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).

[20] Wikipedia contributors. "Precision and Recall". Wikipedia, The Free Encyclopedia. (http://en.wikipedia.org/wiki/Precision_and_recall accessed July 2013).

SN Sausalito News 1885 - 1922	70.4%	17.3%

Raw OCR character versus corrected character accuracy

Title	OCR character accuracy	corrected accuracy
PRP Pacific Rural Press 1871 - 1922	92.6%	99.3%
SFC San Francisco Call 1890 - 1913	92.6%	99.6%
LAH Los Angeles Herald 1873 - 1910	88.7%	99.1%
LH Livermore Herald 1877 - 1899	88.6%	99.9%
DAC Daily Alta California 1841 - 1891	88.2%	99.9%
CFJ California Farmer and Journal of Useful Sciences 1855 - 1880	86.5%	99.8%
SN Sausalito News 1885 - 1922	70.4%	100.0%

Corrected accuracy by newspaper title

Title	OCR character accuracy	~OCR word accuracy*	Corrected accuracy	~Corrected word accuracy
PRP 1871 - 1922	92.6%	68.1%	99.3%	96.5%
SFC 1890 - 1913	92.6%	68.1%	99.6%	98.0%
LAH 1873 - 1910	88.7%	54.9%	99.1%	95.6%
LH 1877 - 1899	88.6%	54.6%	99.9%	99.5%
DAC 1841 - 1891	88.2%	53.4%	99.9%	99.5%
CF 1855 - 1880	86.5%	48.4%	98.3%	91.8%
SN 1885 - 1922	70.4%	17.3%	100.0%	100.0%

Obviously corrected text is far more accurate than raw text, more than 5 times as accurate for the least accurate newspaper title (Sausalito News).

Not surprisingly we see that the aggregate accuracy of corrected text is better than that of uncorrected text. How big is the difference in corrected text accuracy between users? We sampled the corrected text of the 10 most prolific CDNC text correctors. The results are in the nearby table (character accuracy, not word accuracy).

Correction accuracy by user

User	Average OCR accuracy	Correction accuracy
A	70.4%	100.0%
B	87.1%	99.5%
C	95.4%	99.5%
D	86.5%	98.3%
E	95.3%	100.0%
F	91.0%	100.0%
G	91.0%	99.8%
H	90.5%	99.0%
I	96.6%	99.8%
J	94.8%	100.0%
K	86.8%	99.3%

Even the least accurate of the correctors (98.3%) is still far better than raw OCR accuracy. Of course one might suspect that the prolific text correctors are also the most accurate text correctors and less prolific correctors are less accurate. That will be a measurement for another time…

This is a small sample of corrected text and users, but it is reasonable to expect that a larger sample of OCR text and users would give similar results. And anyone with the time and inclination to repeat this measurement with a larger OCR text sample and more users is welcome to the data. Just let us know. ☺

What difference does more accurate text make in practice? Let's again take the surname 'Arndt' and search for it on Chronicling America. A search performed 31 Oct 2012 yielded 10,267 results. Chronicling America has only uncorrected OCR text. If the OCR text has accuracy similar to CDNC's uncorrected OCR text, the accuracy for a 5-character word is 55.8% and there are 8,133 instances of 'Arndt' in Chronicling America that this search did **not** find. On the other hand if Chronicling America had corrected OCR text similar to CDNC's corrected text, the accuracy for a 5-character word is ~97.0% and only 317 instances of 'Arndt' were not found. Quite a difference!

Character accuracy is multiplicative. In other words, for constant character accuracy, longer words have lower accuracy. Here's an example of what this means in practice, again using real data from Chronicling America. We assume raw OCR accuracy (89%) and corrected accuracy (99%) similar to CDNC.

Raw versus corrected text accuracy for words of various lengths

Name	Name length	Raw text accuracy	Corrected text accuracy
Eklund	6	49.7%	94.2%
Kennedy	7	44.2%	93.2%
Espinosa	8	39.4%	92.3%
Bonaparte	9	35.0%	91.4%
Chatterjee	10	31.2%	90.4%

Missing results for raw versus corrected text accuracy

Name	Number of search results	Missing results with raw text accuracy	Missing results with corrected text accuracy
Eklund	2,951	2,987	182
Kennedy	360,723	455,392	26,111
Espinosa	1,918	2,950	160
Bonaparte	44,664	82,947	4,203
Chatterjee	19	42	2

These examples show that genealogists with short family names will fare better in their research on uncorrected OCR text than those with long family names. These calculations are not intended to single out Chronicling America as having particularly egregious OCR text: Any uncorrected digital newspaper collection will have similar accuracies.

4.2 Economic benefit

OCR text correction can be outsourced to service bureaus. Australia, New Zealand, Singapore, CDNC, and others do rely on outsourced OCR text correction, but, 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 40 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 calculation, let's assume USD $0.50 per 1000 characters.

As of July 2013, volunteers at CDNC have corrected 1,273,000 lines of text. Using the assumptions in the preceding paragraph, the value of CDNC volunteer labor is 1,273,000 lines x 40 characters x 1/1000 characters x $0.50 = $25,460. Similar calculations for the National Library of Australia's Trove, where volunteers have corrected 69,918,892 lines of text, values Trove volunteer labor at $2,035,326. For Cambridge Public Library, where volunteers have corrected 43,671 lines of text, the value is $873.

$0.50 per 1000 characters	Lines corrected	Volunteer labor value
Cambridge	43,671	USD $873
CDNC	1,273,000	USD $25,460
Trove	101,766,326	USD $2,035,326

These figures are the estimated monetary value of the avoided labor costs if service bureaus had done the same work as volunteers. But this is not the only way to estimate the value of volunteer labor.

The National Library of Australia reports volunteer labor value by assuming that it takes 15 seconds to correct a line of text and that the volunteers would have been paid the same as the lowest paid Library employee or $37.42[21].

Hourly wage $37.42	Lines corrected	Volunteer labor value
Cambridge	43,671	USD $6,809
CDNC	1,273,000	USD $198,482
Trove	101,766,326	USD $15,867,066

Regardless of which way one chooses to value volunteer labor, the numbers are truly significant[22]! However as we shall see below, the monetary value of volunteer labor may not be crowdsourcing's most significant benefit to cultural heritage digital collections.

5. Crowdsourcing at the California Digital Newspaper Collection

The California Digital Newspaper Collection (CDNC) is the largest, freely accessible archive of digitized California newspapers. The collection contains over 60,000 issues and 550,000 pages—and growing, ranging from 1846 to the present. It is available for searching at http://cdnc.ucr.edu. The project is managed and hosted by the Center for Bibliographical Studies and Research (CBSR) at the University of California, Riverside. It has been supported in part both by the National Digital Newspaper Program (NDNP), a joint effort by the National Endowment for the Humanities and the Library of Congress, and by the Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian. The CDNC has also partnered with local institutions around the state to digitize their newspapers and add the content to the archive.

Work on digitizing California newspapers began in 2005, when the CBSR was selected as one of the first six participants in the NDNP. Much of the initial content for the archive came from data that was also submitted to the NDNP, but all pages were digitized to the article level rather than just the page, a practice the CDNC continues to this day. In October of 2007 the CDNC officially launched its website and in the fall of 2009 began hosting with Veridian software. In August of 2011 the CDNC, working closely with the developers of Veridian, enabled user text correction (UTC) within the archive, allowing users to register and then edit the computer-generated text. In the months since over 1300 individuals have registered, of whom nearly 600 have corrected over a million lines of text.

[21] AUD $40.38 = USD $37.42 (July 2013 exchange rates) This is the actual labor value assumed by the National Library of Australia to calculate avoided costs due to crowdsourced OCR text correction in its 2012 Trove Status Report**.**

[22] If you don't like the assumptions, put numbers you like into one of the following formulas: *linesCorrected x charactersPerLine / 1000 x costPer1000Characters* or *linesCorrected x 15sec x 1/3600 x hourlyWage.*

The CDNC has maintained Google Analytics ever since the project started using Veridian and the statistics for average visit duration reveal the impact of UTC on the archive. In the fall of 2010 duration dropped precipitously from 16:28 to 8:30 minutes; time spent on the CDNC declined, not surprisingly, as indexing by Google and the bounce rate increased with the transition to Veridian. Between November 2010 and July 2011 the average visit duration fluctuated but never went above 10:41 minutes, averaging 9:14 minutes during those nine months. Then in August 2011, when UTC was introduced, the average duration jumped to 11:52 minutes. For the next year it remained fairly consistent, averaging 10:42 minutes. In other words, adding UTC to the site increased the average monthly visit duration by over one minute, a number that is particularly impressive when one considers that both the number of visitors per month and the bounce rate continued to increase during the same period.

In June of 2013 the CDNC added user tagging and comments. As of August 1, despite limited announcement of the new features, users had contributed over 100 tags. We expect tagging to increase substantially in the coming months and plan to publicize the feature more aggressively. One interesting development so far is that users have created tags that could be linked to other resources. For example, there are already two tags for Abdu-l Baha, the founder of the Bahia Faith. Those tags could be linked to the entries in Wikipedia and the VIAF, among other resources, to provide other users more information on the name. We are exploring how best to link out to other authority databases. Stay tuned.

6. Citizen Archivists and Historic Newspaper Collections at the Cambridge Public Library

The concept of citizen archivists is new to libraries and archives and is quickly becoming a trend across the globe. Citizen *archivy*[23] engages the public with archival collections by asking anyone with an Internet connection to enhance or add to existing online historical collections. The beginnings of the citizen archivist can be traced to the National Library of Australia, when in 2009, it created Trove - an interactive search engine of Australian history - and has come full circle with the National Archives and Records Administration (NARA) unveiling this past March a citizen archivist dashboard, where anyone can transcribe text, tag photos, and upload images.

The Cambridge Public Library's Archives and Special Collections has also begun to build digital collections that ask for citizen archivists' help. In March 2013, the library celebrated its one-year anniversary of the launch of the Historic Cambridge Newspaper Collection (http://cambridge.dlconsulting.com), its first digital project that interacts with the user.

Working with DL Consulting, a team of software engineers and system administrators based in New Zealand, the library digitized and made available all Cambridge newspapers free from copyright, including the *Cambridge Chronicle* - the oldest, continually published weekly in the United States. Using Veridian, DL Consulting's newspaper database software platform that encourages and allows users to correct garbled text created during the digitization process, the library asks the public to directly engage with the historic materials and improve the resource for the greater good. The collection tracks the number of lines each user has corrected and lists those who have corrected the most lines prominently on the homepage in a "Text Corrector Hall of Fame."

[23] *Archivy* is defined as the discipline of archives.

Interesting text correction trends have emerged: From those who correct just a few lines to text correcting standoffs for the top spot in the Hall of Fame. After one year, users have corrected over 40,000 lines of unreadable newspaper text, and as word spreads and more people are working with the collection, the amount of corrections has grown exponentially – up to 1,000 lines of text per week. These increasing numbers are impressive considering that the library has not yet marketed the collection or the idea of involving users on the level of citizen archivists.

Despite the soft launch of the collection, people from all over the world are searching, reading, and enhancing Cambridge's historic newspapers, which previously could only be viewed on microfilm at the library. Having several interested people help improve the search capabilities of the collection by correcting text has the advantage of both creating a community of users who are excited about revealing local history as well as solving the very real problem of staff time and cost that so many libraries face. The success of this project shows that the public has a real desire to be involved at a local level preserving and making available historical resources. Anyone can be a citizen archivist if given the opportunity. Building on this success, the library will continue to make available digital collections that empower users by allowing them to create and improve content, build virtual communities around historical and genealogical research, and, best of all, make it an enjoyable experience that keeps them engaged.

7. Real benefit of crowdsourcing

Improved search is of course a real benefit and one that is easily measured. But, like Clay Shirky in *Cognitive Surplus*, Trevor Owens in his blog argues that the most important benefit of crowdsourcing cultural heritage collections is the meaningful activity and the facility for purposeful contributions that it provides for volunteers. Here are 2 excerpts from his excellent blog on the objectives of crowdsourcing[24]:

> *What crowdsourcing does, that most digital collection platforms fail to do, is 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. Far from being an instrument which enables us to ultimately 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, 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.*

These benefits, unlike improved search accuracy or avoided costs of digitization, are not easily measured. But although they are intangible and difficult-to-quantify, they are

[24] 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/

nonetheless very real. How does one value the opportunity to contribute to the public memory through correcting OCR text correction, transcribing diaries from the War Between the States, or entering data from census records? As we saw from their comments above, CDNC, Trove, and Cambridge Public Library volunteers have each found personal value in their text correction. And regardless of the difficulty to measure or quantify, crowdsourcing in a very simple way increases libraries' relevance to the communities they serve in the age of the Internet.

Appendix: **User Survey for California Digital Newspapers Collection**

The California Digital Newspaper Collection (CDNC) has been online since 2007. The purpose of this survey is to learn about the users of the website, for what purposes the collection is used, and to discover how many users correct text. We hope that you will complete the survey and become a regular at CDNC. Watch for additional new features to the site in the near future.

1. Do you use CDNC for work-related research or for personal purposes or for both?
☐ Work
☐ Personal
☐ Both
☐ Don't use the collection

2. Do you consider yourself a genealogist or family historian?
☐ Yes
☐ No

3. What are the main reasons you use CDNC (check all that apply)?
☐ Genealogy or family history research
☐ Community history
☐ California history
☐ Regional history of the West
☐ General historical research
☐ Other (please specify)

4. What type of information do you search for (check all the apply)?
☐ Birth announcements
☐ Wedding announcements
☐ Death announcements and obituaries
☐ Biographical information
☐ General community history
☐ Legal or court notices
☐ Advertisements
☐ Other (please specify)

5. Do you participate in any online genealogy forums?
☐ FamilySearch
☐ Ancestry.com
☐ GenForum
☐ MyHeritage
☐ RootsChat
☐ RootsWeb
☐ Do not participate in any forums
☐ Other (please specify)

6. Approximately how often do you visit CDNC?
□ Daily
□ Weekly
□ Monthly
□ Never

7. For a typical visit, estimate the number of minutes you spend on the CDNC website?
____ Minutes

Text at the CDNC website is computer generated using optical character recognition (OCR). The tet has many errors and consequently search results are less than perfect.

The website has a text correction feature that anyone can use. Corrected text is searchable by other users and gradually, through user contributions, the accuracy for website users will improve.

8. Do you currently use CDNC's text correction?
□ Yes
□ No

9. If you didn't know about CDNC text correction before taking this survey, will you try text correction in future?
□ Yes
□ No
□ Maybe

10. Do you have a social networking account?
□ No social network account
□ Delicious
□ Facebook
□ Google+
□ Twitter
□ Other (please specify)

11. Have you ever shared an artice or information you found in CDNC via social media such as Twitter or Facebook?
□ Yes
□ No

If you answered yes to the previous question, please briefly explain what you shared and which social media account you used. _____

12. We hope to add new features to the Veridian software used to host the CDNC. Which of the following features would you be likely to use? (Answer with **0** for wouldn't use, answer with **1** for might use, or answer with **2** for would definitely use.)

____ Download results from a search in spreadsheet of database format
____ Apply "tags" to an article for future recovery or clarification of its content

____ Write comments about an article
____ Download a high quality image of a page
____ Interact with other users on a CDNC forum
____ Save search results, articles, or pages to a personal collection
____ Add outside references to an article, for example, from Wikipedia or another website
____ Add a "georeference" to an article by linking it to a place on a map allowing users to see collections of articles about a specific place
____ Describe any other features you would like to be added to CDNC

13. Please give some basic demographic information about yourself.
____ State / Province
____ ZIP / Postal Code
____ Country

14. What is your age?
☐ Under 20
☐ 20 to 29
☐ 30 to 39
☐ 40 to 49
☐ 50 to 59
☐ 60 to 69
☐ 70+

If you would not mind being contacted for further questions about your use of CDNC and about improvements or additional features that you would like, please give your name and email address and/or telephone number. This information will be held in strictest confidentiality and will by used by CDNC only for the aforementioned purposes. And of course you are always welcome to contact CDNC with questions and requests at cdnc@cbsr.ucr.edu.

_____ Name
_____ Country
_____ Email address
_____ Phone number