Crowdsourcing Cultural Heritage

Alyssa Pacy
Archivist
Cambridge Public Library

Digital Commonwealth Conference
April 8, 2014

Presentation based on a paper by Frederick Zarndt, Brian Geiger, Alyssa Pacy, and Stefan Boddie, “Crowdsourcing the World’s Cultural Heritage, Part II.”

Crowdsourcing

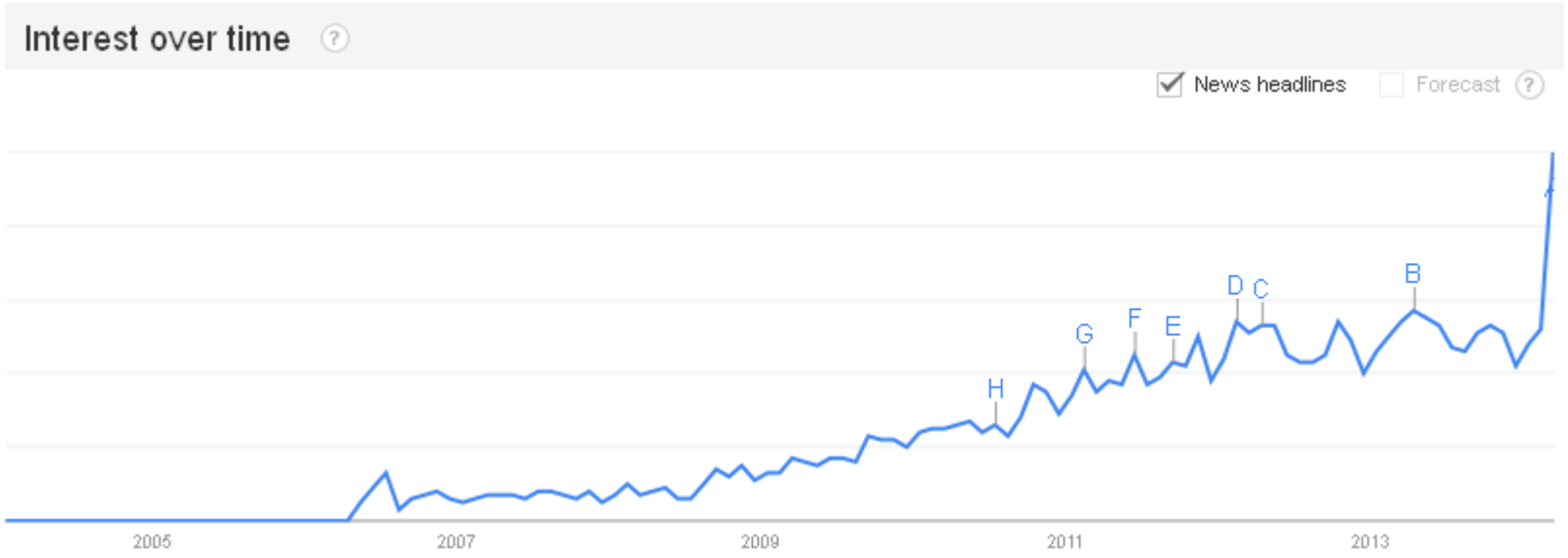
Jeff Howe, who coined the word “crowdsourcing” drafted 4 principles describing the new labor pool:

The crowd is

1. dispersed
2. a short attention span
3. full of specialists
4. finds the best stuff

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

Crowdsourcing



Google Trends "crowdsourcing" 2004 to March 2014

Crowdsourcing

"**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."

Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. "Towards an integrated crowdsourcing definition." Journal of Information Science. 2012. pp. 1-14.

Historic Cambridge Newspaper Collection

Newspapers

Cambridge Chronicle, 1846-1923
Cambridge Press, 1887-1889
Cambridge Sentinel, 1903-1912
Cambridge Tribune, 1887-1923

Collection Statistics

6,346 issues comprising 59,070 pages, and 669,406 articles

40,634 subject cards and 13,099 obituary cards that index Cambridge newspapers published between 1950 and 2008.

Historic Cambridge Newspaper Collection

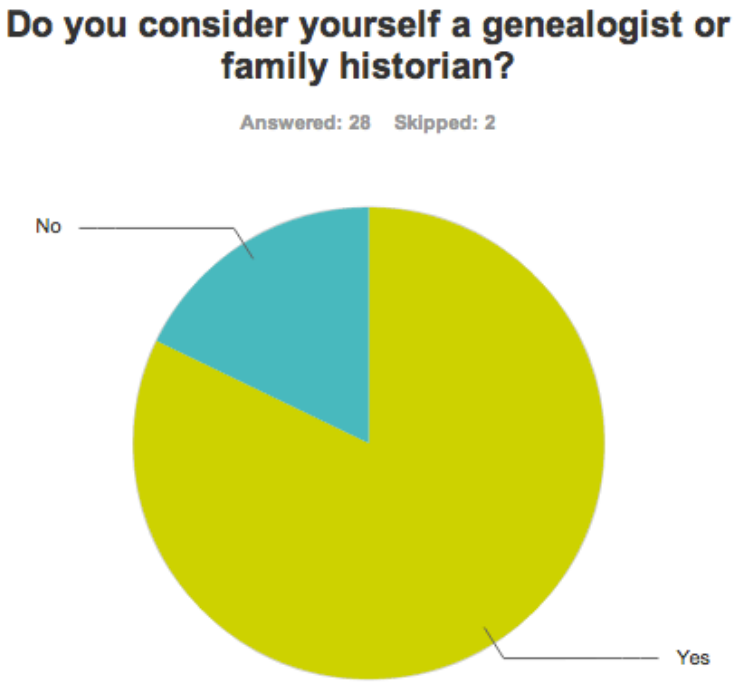
Text Correction Statistics

Number of Registered users: 192

Number of users who have corrected text: 98

Total number of lines of text corrected: 153,126

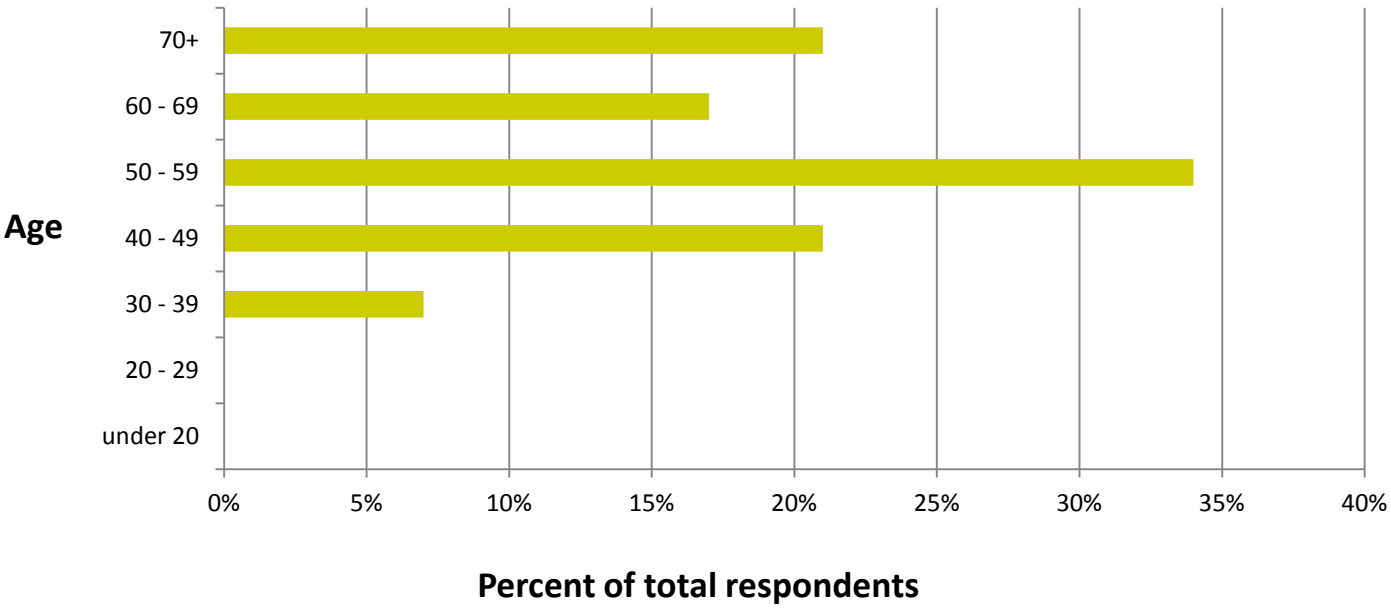
Demographics



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

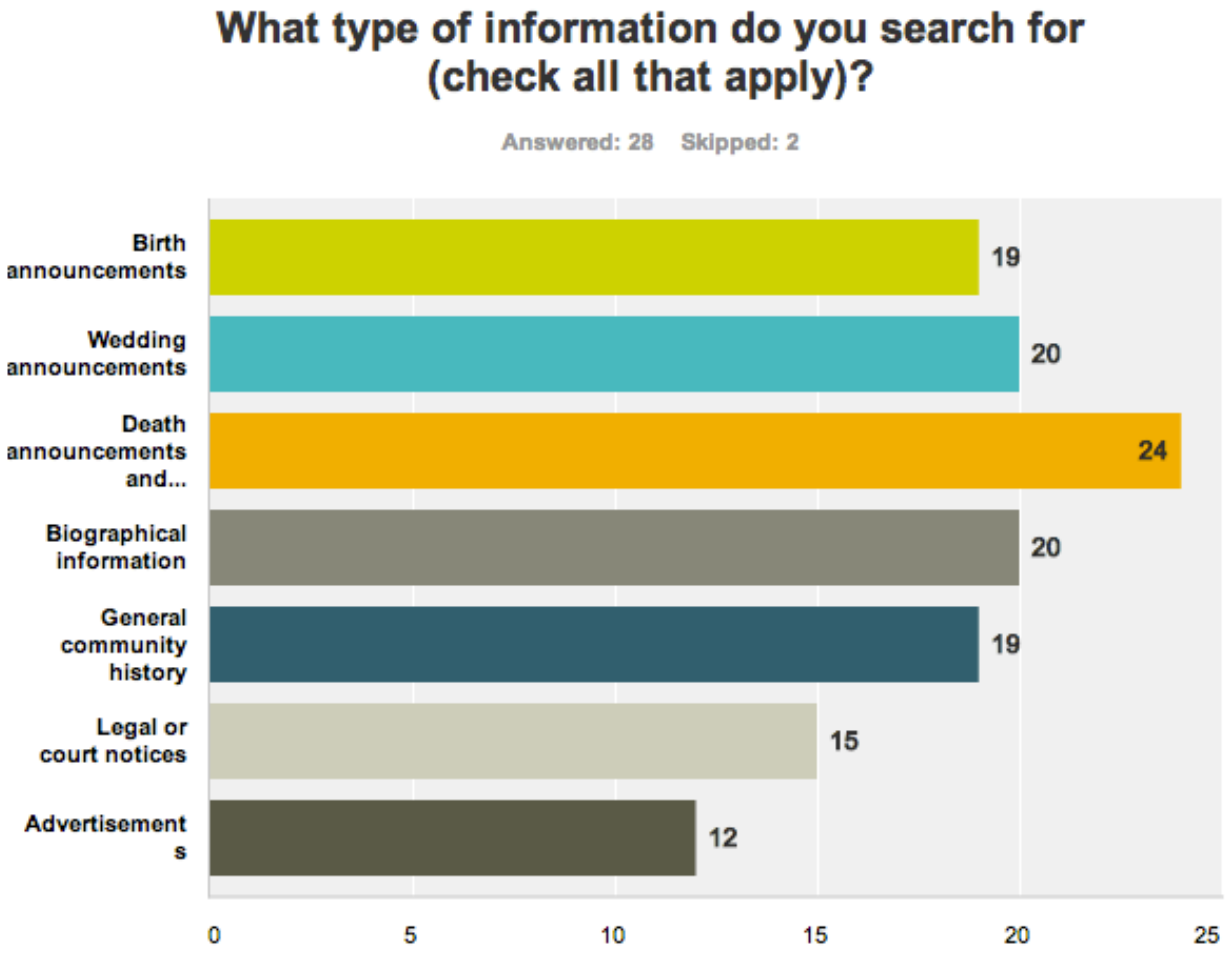
Survey of users of the Cambridge Historic Newspaper Database, February to July, 2013.

Demographics



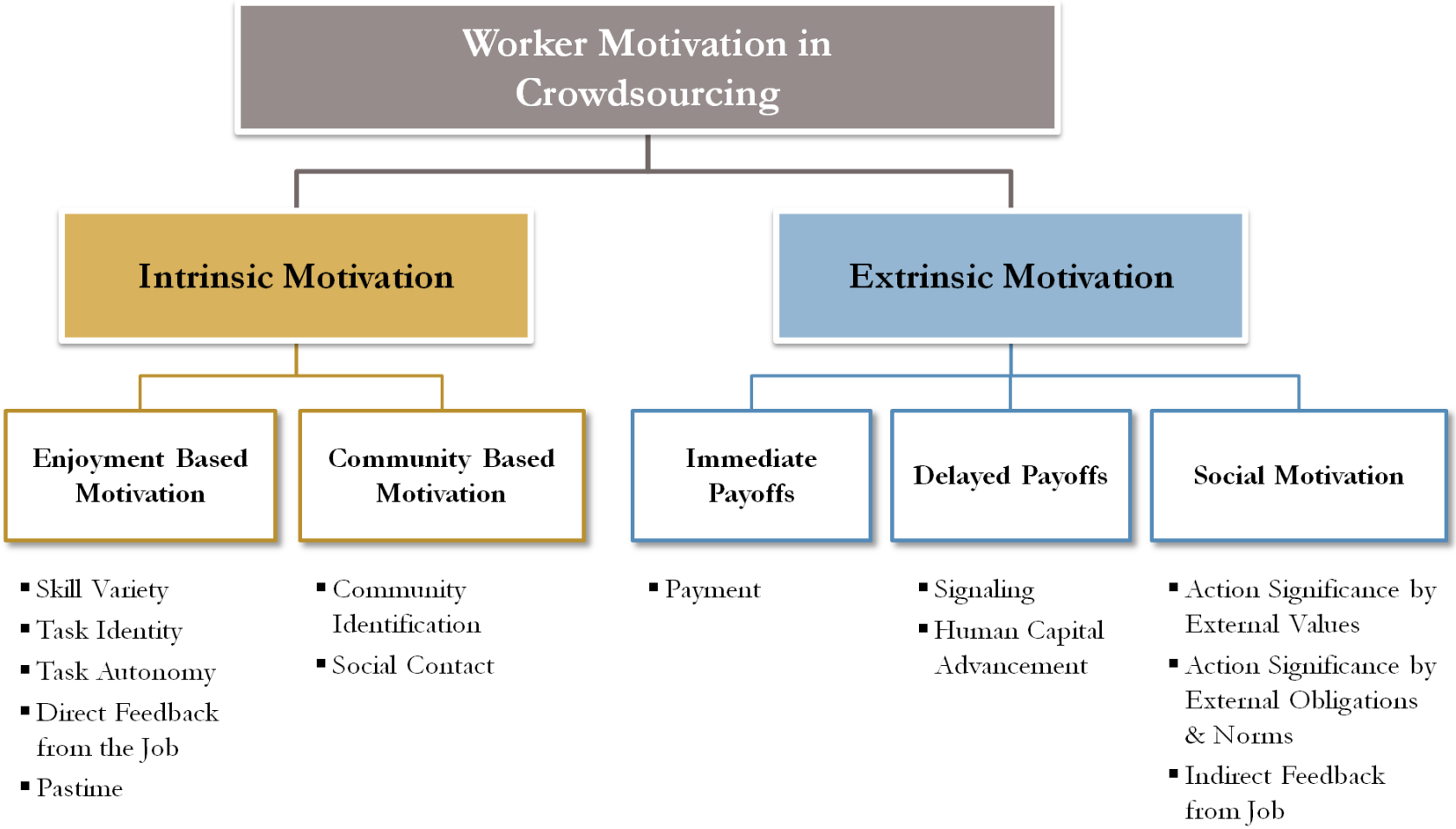
Survey of users of the Cambridge Historic Newspaper Database, February to July, 2013.

Demographics



Survey of users of the Cambridge Historic Newspaper Database, February to July, 2013.

Motivation



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.

Motivation

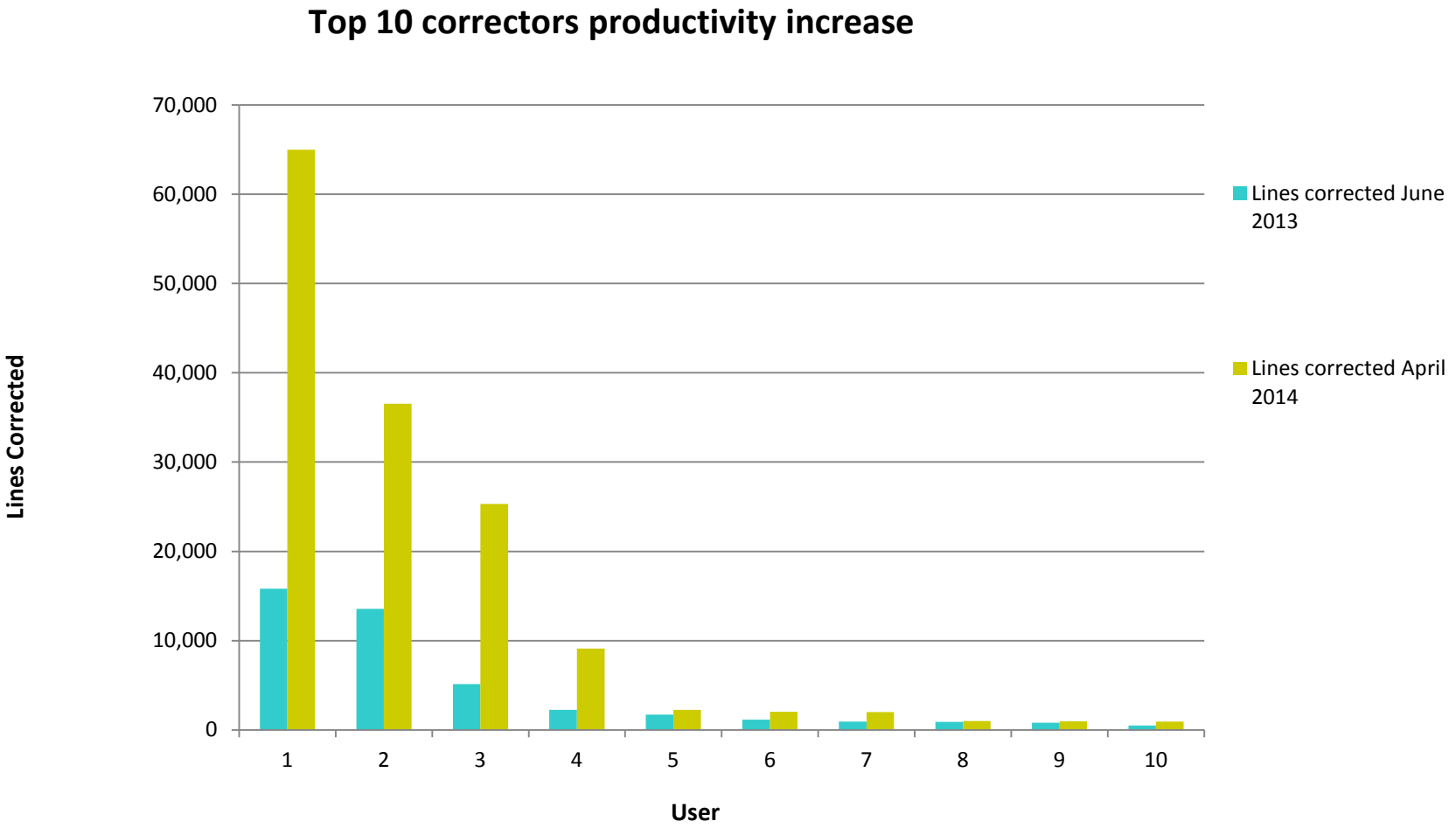
"As an amateur historical researcher my time for research is very limited. Making time to travel to archives, libraries, and historical societies does not happen as often as I would like. The Cambridge Public Library's online newspaper collection has been an invaluable resource and it is fun. I am very 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

Motivation

Lines Corrected	User
65,000	1
36,544	2
25,321	3
9,101	4
2,252	5
2,032	6
2004	7
983	8
957	9
922	10

Motivation


Top 10 correctors productivity increase
Lines Corrected
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
1
2
3
4
5
6
7
8
9
10
User
Lines corrected June 2013
Lines corrected April 2014

Benefits: Improved Text Accuracy

Raw OCR Accuracy

20th Century Newspapers = 68%

Sampling of Australia's newspaper (1803 to 1954) = 71%-98%

Raw Word Accuracy

With 90% OCR accuracy, 60% or 6 words out of 10 are correct

Longer words have lower accuracy

i.e., "Arndt" will get more hits than "Chatterjee"

Benefits: Improved Text Accuracy

CNDC Correction accuracy by user Character accuracy		
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%

Benefits: Cost and Labor

Cost of outsourced OCR text correction: $.50 per 1000 characters

Fiscal Year	Lines corrected	Volunteer labor value
2012	10,894	USD$217.88
2013	27,313	USD$546.26
2014	114,919	USD$2,298.38

Assuming that:
Outsourced text correction has 99.5% accuracy
and
An average newspaper has 40 characters per line

Calculation:
Lines corrected x characters per line x 1/1000 characters x $0.50

Benefits: Cost and Labor

Cost of outsourced OCR text correction: $.50 per 1000 characters

	Lines corrected	Volunteer labor value
Cambridge	153,126	USD$3,062.52
CNDC	2,369,497	USD$47,389.94
Trove	101,766,326*	USD$2,035,326.52

*This number is from June 2013.

Benefits: Cost and Labor

Hourly wage $14.71*	Lines corrected	Volunteer labor value
Cambridge	153,126	USD$9,382.72
CNDC	2,369,497	USD$145,189.75
Trove	101,766,326	USD$6,235,681.25

*$14.71 is the minimum hourly wage of a Cambridge Public Library employee.

Assuming that:
It takes 15 seconds to correct a line of text

Calculation:
linesCorrected x charactersPerLine / 1000 x costPer1000Characters
or
linesCorrected x 15sec x 1/3600 x hourlyWage.

The Real Benefit of Crowdsourcing Cultural Heritage

*"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**."*

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/

Summary

- Citizen Archivists and Patron Curators
- Creates a community of users
- Fulfills the public's desire to be involved at a local level preserving and making available historical resources
- Personal value in volunteers correcting text
- Serendipity of discovery while correcting text
- Increases libraries' relevance to the communities they serve in the age of the Internet