Putting the world’s cultural heritage online with crowdsourcing

Frederick Zarndt
sponsored by
CCS / Digital Divide Data / DL Consulting
Crowds
In 2004 James Surowiecki published “The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations”. In it he asserts

a crowd of persons that are diverse, independent, and decentralized usually make better judgements or decisions than single persons
“crowdsourcing”

was coined by Jeff Howe in “The rise of crowdsourcing” published in Wired magazine June 2006.
A Google advanced search for “crowdsourcing” from 1-Jun-2006, the date of publication of Jeff Howe’s Wired magazine article, to 1-Jun-2007 gives 44,600 hits.

A date range of 1-Jun-2011 to 1-Jun-2012 gives 2,680,000 hits.
The Crowdsourcing Process
In Eight Steps

1. Company has a problem
2. Company broadcasts problem online
3. Online "crowd" is asked to give solutions
4. Crowd submits solutions
5. Crowd vets solutions
6. Company rewards winning solvers
7. Company owns winning solutions
8. Company profits

Image by Daren C. Brabham | www.darenbrabham.com
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.

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.
Amazon Mechanical Turk was launched Nov 2005
Alexa global rank of Amazon Mechanical Turk (13-Jun-2012): 6,022
Welcome to Galaxy Zoo, where you can help astronomers explore the Universe

Galaxy Zoo: Hubble uses gorgeous imagery of hundreds of thousands of galaxies drawn from NASA's Hubble Space Telescope archive. To understand how these galaxies, and our own, formed we need your help to classify them according to their shapes — a task at which your brain is better than even the most advanced computer. If you're quick, you may even be the first person in history to see each of the galaxies you're asked to classify.

More than 250,000 people have taken part in Galaxy Zoo so far, producing a wealth of valuable data and sending telescopes on Earth and in space chasing after their discoveries. The images used in Galaxy Zoo: Hubble are more detailed and beautiful than ever, and will allow us to look deeper into the Universe than ever before. To begin exploring, click the 'How To Take Part' link above, or read The Story So Far to find out what Galaxy Zoo has achieved to date.

Thanks for your help, and happy classifying.

The Galaxy Zoo team.
Wikipedia:About

From Wikipedia, the free encyclopedia

This is a general introduction for visitors to Wikipedia. The project also has an encyclopedia article about itself, Wikipedia, and some introductions for aspiring contributors.

For Wikipedia's formal organizational structure, see Wikipedia:Formal organization.

For Wikipedia namespace, see Wikipedia:Project namespace.

For help aimed at readers only, see Help:About.

Wikipedia ([wikipedia.org](http://wikipedia.org) or [wikipedia.org/wiki](http://wikipedia.org/wiki)) is a multilingual, web-based, free-content encyclopedia project based on an open publishing model. The name "Wikipedia" is a portmanteau of the words wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning "quick") and encyclopedia. Wikipedia's articles provide links to guide the user to related pages with additional information.

Wikipedia is written collaboratively by largely anonymous Internet volunteers who write without pay. Anyone with Internet access can read and make changes to Wikipedia articles (except in certain cases where editing is restricted to prevent disruption or vandalism). Users can contribute anonymously, under a pseudonym, or with their real identity, if they choose.

The fundamental principles by which Wikipedia operates are the Five pillars. The Wikipedia community has developed many policies and guidelines to improve the encyclopedia; however, it is not a formal requirement to be familiar with them before contributing.

Since its creation in 2001, Wikipedia has grown rapidly into one of the largest reference websites, attracting 400 million unique visitors monthly as of March 2011 according to ComScore.[1] There are more than 85,000 active contributors working on more than 21,000,000 articles in more than 280 languages. As of today, there are 3,969,640 articles in English. Every day, hundreds of thousands of visitors from around the world collectively make tens of thousands of edits and create thousands of new articles to augment the knowledge held by the Wikipedia encyclopedia (see also Wikipedia:Statistics.)

People of all ages, cultures and backgrounds can add or edit article prose, references, images and other media here. What is contributed is more important than the expertise or qualifications of the contributor. What will remain depends upon whether it fits within Wikipedia's policies, including being verifiable against a published reliable source, so excluding editors' opinions and beliefs and unreviewed research, and is free of copyright restrictions and contentious material about living people. Contributions cannot damage Wikipedia because the software allows easy reversal of mistakes and many experienced editors are watching to help and ensure that edits are cumulative improvements. Begin by simply clicking the edit link at the top of any editable page!

Wikipedia is a live collaboration differing from paper-based reference sources in important ways. Unlike printed encyclopedias, Wikipedia is continually created and updated, with articles on historic events appearing within minutes, rather than months or years. Older articles tend to grow more comprehensive and balanced; newer articles may contain misinformation, unencyclopedic content, or vandalism. Awareness of this aids obtaining valid information and avoiding recently added misinformation (see Researching with Wikipedia).

What Wikipedia is not circumscribes Wikipedia's scope. Further information on key topics appears below. Further advice is at Frequently asked questions, advice for parents, or see Where to ask questions. For help getting started with editing or other issues, see Help:Contents.
Wikipedia

- Began 2001
- Now in 285 languages
- 3,900,000+ articles in English, 1,400,000+ in German, 1,250,000+ in French, 1,050,000 in Dutch
- 40 wikipedia languages with more than 100,000 articles
- 112 wikipedia languages with more than 10,000 articles
- 400,000,000 unique visitors per month
- 85,000 active contributors
- Ranked #6 in worldwide web traffic
About Us

Transcribe Bentham is an award-winning participatory project based at 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. We would like to encourage all those who have an interest in Bentham or those with an interest in history, politics, law, philosophy and economics, fields to which Bentham made significant contributions, to visit the site. Those with an enthusiasm for palaeography, transcription and manuscript studies will be interested in Bentham’s handwriting, while those involved in digital humanities, education and heritage learning will find the site intriguing. Undergraduates and school pupils studying Bentham’s ideas are particularly encouraged to use the site to enhance their learning experience.

Bentham Papers

There are 60,000 papers written by Bentham in UCL’s library but several thousands of these papers, potentially of immense historical and philosophical importance, have yet to be transcribed and studied. By transcribing this material for the first time, you will be doing two important tasks:

- making Bentham’s thought accessible to the world at large
- helping UCL’s Bentham Project, which was founded in 1959 to produce the new, authoritative edition of the Collected Works of Jeremy Bentham
Family Search Indexing was 1st launched (beta) 2004
Alexa global traffic rank of FamilySearch (13-Jun-2012): 4,419
• Started (beta) 2004

• More than 780,000 worldwide registered volunteers from ~25 countries index records relevant to family history

• Approximately 100,000 active volunteers each month

• UI in Chinese, English, German, French, Italian, Japanese, Korean, Portuguese, and Russian

• Blind double-key entry with arbitration / reconciliation

• More than 1,500,088 records indexed (July 2012)

• Accuracy typically > 99.95%
Project Gutenberg was 1st launched Dec 1971
Alexa global traffic rank of Project Gutenberg (13-Jun-2012): 5,744
• Started Dec 1971

• Worldwide volunteers transcribe or proofread OCR’d public domain books through Distributed Proofreaders

• 40,000 books (July 2012)

• Partner / affiliated projects for Australia, Canada, Europe, Germany, Luxembourg, Philippines, Runeberg (Nordic literature), Russia, Taiwan
National Library of Australia

- Online since 2008
- 7,200,000+ pages
- Top text corrector 1,250,000 lines (June 2012)
- 2,450,000+ lines corrected each month (1st 6 months 2012)
- 68,908,757 lines corrected as of July 2012, up from 42,411,468 lines corrected July 2011.
- 63,613 total registered users (July 2012)
- 4,146 active users (June 2012)
FEATURED

Sausalito News 22 March 1888

SEARCH

ABOUT

This collection contains 55,970 issues comprising 495,175 pages and 5,658,224 articles.

The California Digital Newspaper Collection is a project of the Center for Bibliographical Studies and Research (CBSR) at the University of California, Riverside.

The CDNC is supported in part by the U.S. Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, administered in California by the State Librarian.

The CBSR has received three grants from the National Endowment for the Humanities to digitize California newspapers for the National Digital Newspaper Program. Titles digitized as part of the NDNP are available both here and at the Library of Congress Chronicling America website.

We are eager to know what users think of this site. Please email your comments to cbsinfo@ucr.edu.

Like the CDNC on Facebook.

DONATE

Though access to the CDNC is free, maintaining and improving it is not. Please consider supporting the CDNC.

TOP TEXT CORRECTORS

1. Wes Keat 155817
2. anhh 59815
3. daveg 18898
4. Sarah Draper 17513
5. chariq 16875

More information...
California Digital Newspaper Collection

• CDNC began digitizing newspapers in 2005 as part of NDNP
• Hosted on Veridian beginning 2009
• Currently ~500,000 pages
• User OCR correction added August 2011
• ~395,000 lines of text corrected (July 2012)
• Top corrector 155,000 lines > 2x 2nd corrector
Worker Motivation in Crowdsourcing

Intrinsic Motivation
- Enjoyment Based Motivation
  - Skill Variety
  - Task Identity
  - Task Autonomy
  - Direct Feedback from the Job
  - Pastime
- Community Based Motivation
  - Community Identification
  - Social Contact

Extrinsic Motivation
- Immediate Payoffs
  - Payment
- Delayed Payoffs
  - Signaling
  - Human Capital Advancement
- Social Motivation
  - Action Significance by External Values
  - Action Significance by External Obligations & Norms
  - Indirect Feedback from Job

Motivation

Graphic from Kaufmann et al. “More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk.”
Motivation

Genealogists and family historians

- **Trove**: National Library of Australia guesses that ~80% of Trove digitized newspapers users are family historians.

- **PapersPast**: National Library of New Zealand survey found that ~50% of PapersPast users are genealogists.

- **CDNC**: California Digital Newspaper Collection survey found that ~70% of its users are genealogists; 75% are 50 years old or older.
Motivation

Trove users’ 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.”

• “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.”

From Rose Holley in “Many Hands Make Light Work.”
Motivation

CDNC users’ report

• “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.”

Personal communication with CDNC text correctors.
Motivation

“when someone transcribes a document, they are actually better fulfilling the mission of a cultural heritage organization than someone who simply stops by to flip through the pages”

From Trevor Owen’s Crowdstorming blog http://crowdstorming.wordpress.com/
Motivation

“in addition to increasing search accuracy or lowering the costs of document transcription, crowdsourcing is the single greatest advancement in getting people using and interacting with library collections”

Paraphrased from Trevor Owen’s Crowdstorming blog
http://crowdstorming.wordpress.com/
Website traffic

Select an Item to Begin Transcribing:

- Turner S. Bailey diary, 1863
- Philip H. Conard diary, 1864-1865
- Joseph Franklin Culver papers, Mar 1860-Dec 1862
- Joseph Franklin Culver papers, 1863
- Joseph Franklin Culver papers, Jan.-Apr. 1864
- Joseph Franklin Culver papers, May-Nov. 1864
- Joseph Franklin Culver papers, Jan.-June 1865
- William Titus Rigby letters, Feb.-Apr. 1863
- William Titus Rigby letters, June-Dec. 1864
- William Titus Rigby letters, 1865-1868
- William Titus Rigby letters, 1860s-1910s
- Shelton family letters, 1864-1936
- James B. Weaver letters, 1856-1858
- James B. Weaver letters, 1860-1864

Current Progress:
As of July 12, 2012 there have been 14186 pages transcribed.

University of Iowa Libraries
Civil War Diaries & Letters Digital Collection
Civil War Diaries & Letters Transcription Project

Rediscovering Voices
Thanks to the development of "crowdsourcing" or collaborative transcription of manuscript materials, libraries are now able to use the knowledge and interest of the general public to meet goals that they would never have had the time, financial, and staff resources to achieve on their own. Please help us improve access by transcribing the hand-written pages in this collection.
After a crowdsourcing transcription project of diaries from the American War Between the States, Nicole Saylor, Head of Digital Library Services at the University of Iowa Libraries, reported

“On June 9, 2011, we went from about 1000 daily hits to our digital library on a really good day to more than 70,000.”
Website traffic

Changes in website traffic at CDNC after implementing crowdsourcing were not so dramatic as for the University of Iowa Libraries

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>visits</td>
<td>16,934</td>
<td>20,758</td>
<td>+22.6%</td>
</tr>
<tr>
<td>unique visitors</td>
<td>11,030</td>
<td>12,951</td>
<td>+17.4%</td>
</tr>
<tr>
<td>visit duration</td>
<td>9m 24s</td>
<td>11m 6s</td>
<td>+18.1%</td>
</tr>
<tr>
<td>bounce rate</td>
<td>51.3%</td>
<td>44.7%</td>
<td>-12.9%</td>
</tr>
</tbody>
</table>
Crowdsourcing benefits
Financial value of OCR text correction?

Assumptions

• 25 to 50 characters per line in a newspaper column: Assume 35 characters per line

• Outsourced text transcription or correction costs USD $0.35 to $1.20 per 1000 characters: Assume $0.50 per 1000 characters
Economics

- CDNC: 394,365 lines × 35 characters per line × 1/1000 × $0.50 = $6,901 $$
- Trove: 69,918,892 lines × 35 characters per line × 1/1000 × $0.50 = $1,223,581 $$$$$
Text accuracy

- Edwin Kiljin (Koninklijke Bibliotheek the Netherlands) reports raw OCR character accuracies of 68% for early 20th century newspapers

- Rose Holley (National Library of Australia) reports raw OCR character accuracy varied from 71% to 98% on a sample Trove digitized newspapers


Optimistically assume that average raw OCR character accuracy is 90%.

Average length of an English word is 5 characters.

Average word accuracy is $90\% \times 90\% \times 90\% \times 90\% \times 90\% = 59\%$ (6 out of 10 words correct).
Search recall
no text correction

instances of “ARNDT” found

instances of “ARNDT” not found
Assume the crowd corrects OCR text to 99.5% accuracy.

Average word accuracy is now 99.5% x 99.5% x 99.5% x 99.5% = 97.5% (6 out of 10 words correct).
Search recall with text correction

instances of “ARNDT” found

instances of “ARNDT” not found