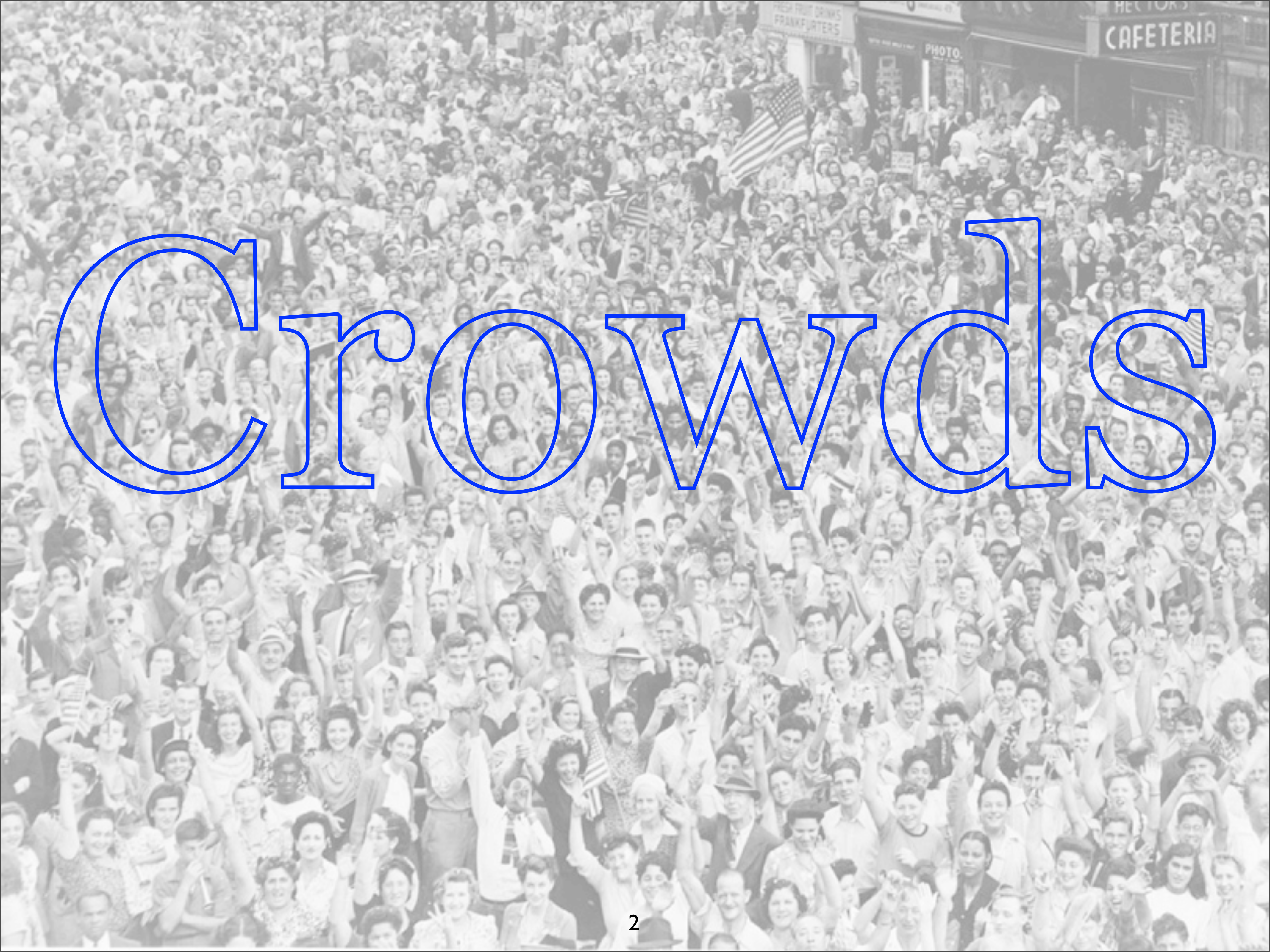
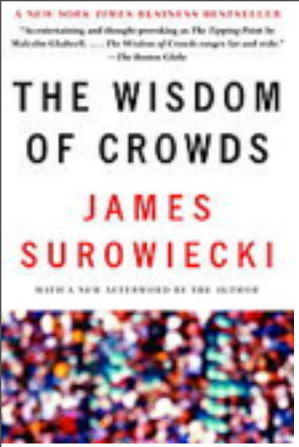
Putting the world's cultural heritage online with crowdsourcing

Frederick Zarndt
sponsored by
CCS / Digital Divide Data / DL Consulting

Crowds





The Wisdom of 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*

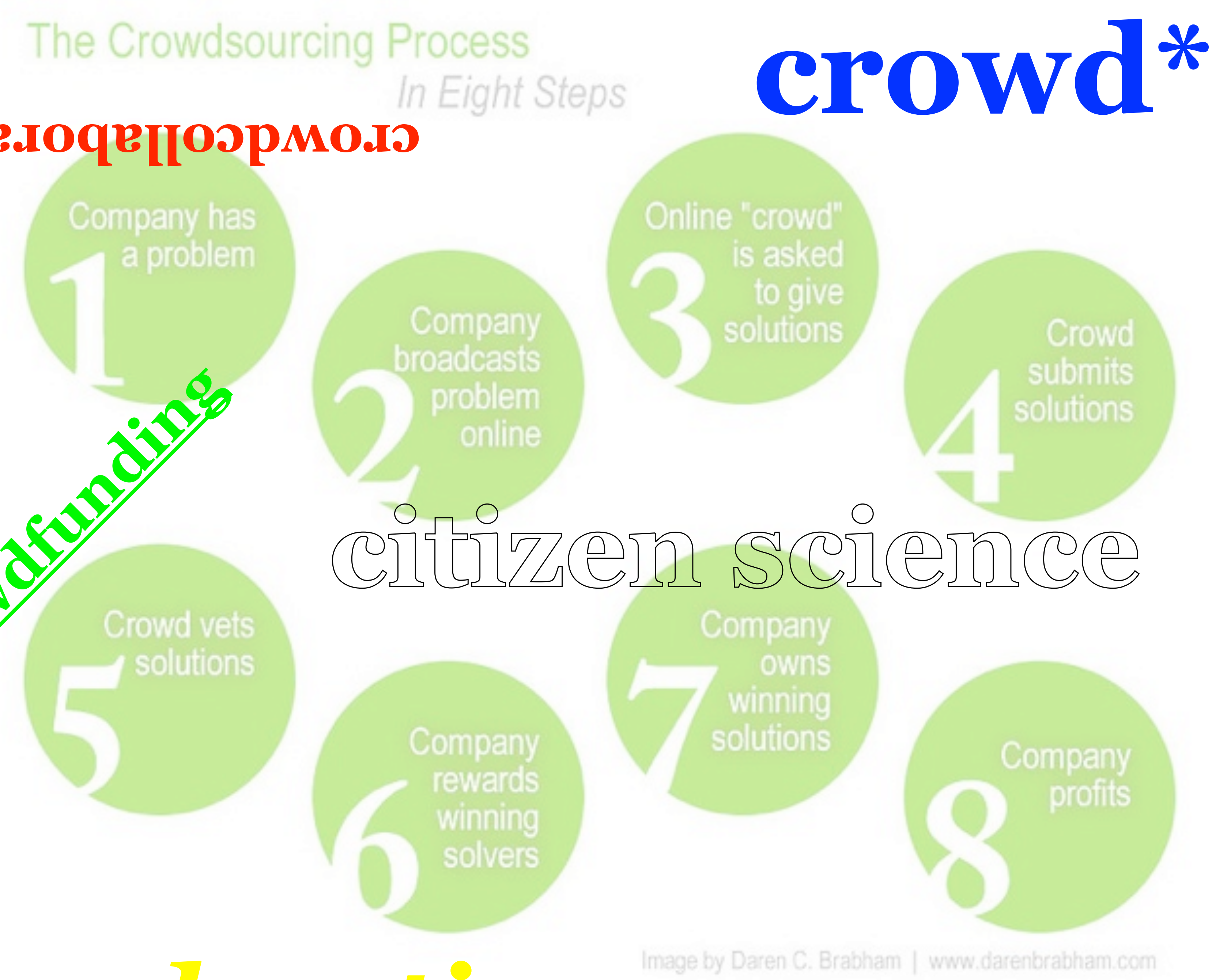
“**crowdsourcing**”

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.

crowd*



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

crowdcollaboration

crowdsourcing

crowdfunding

citizen science

crowdcasting

crowdvoting

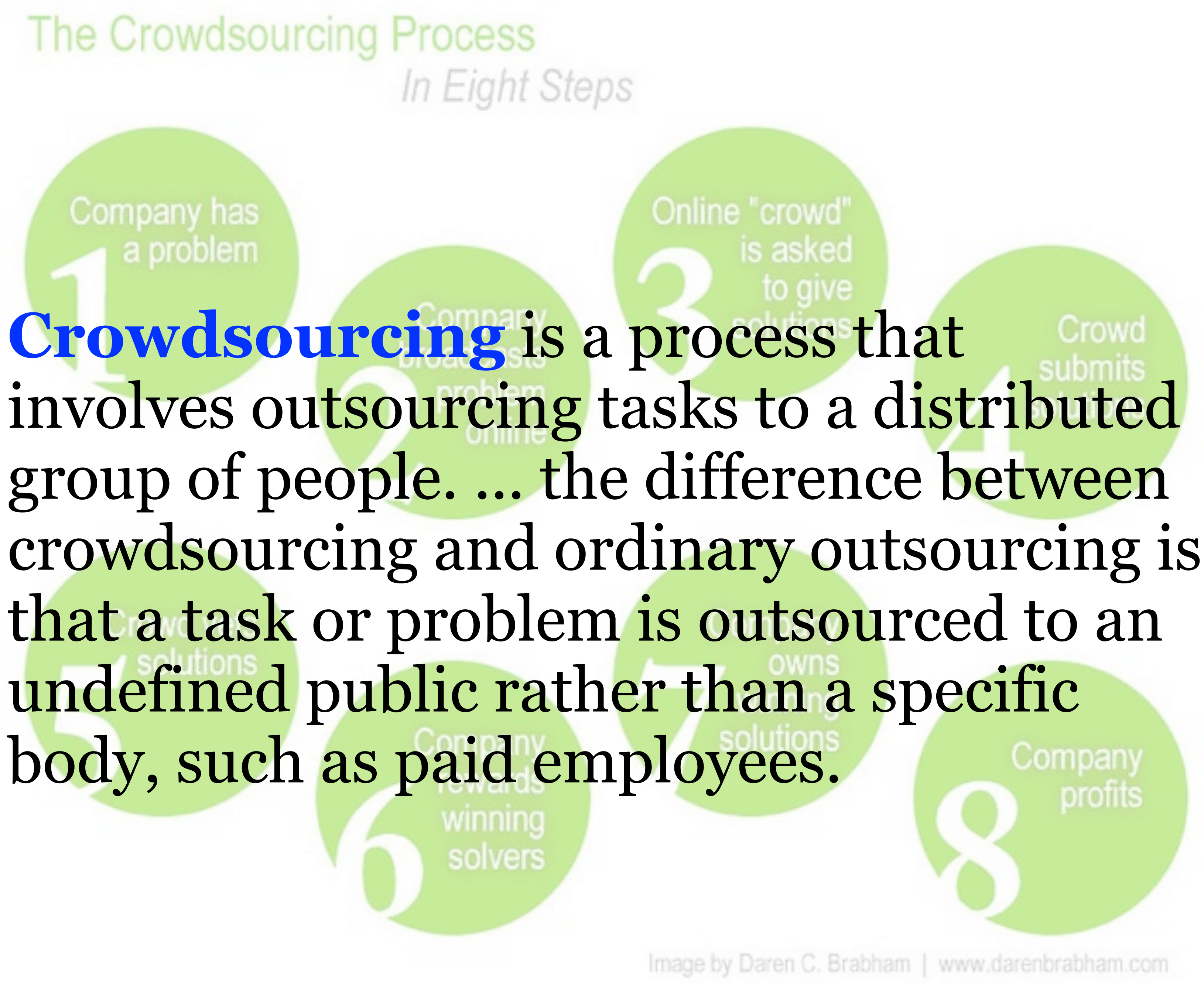

The Crowdsourcing Process
In Eight Steps
1
Company has a problem
2
Company broadcasts problem online
3
Online "crowd" is asked to give solutions
4
Crowd submits solutions
5
Crowd vets solutions
6
Company rewards winning solvers
7
Company owns winning solutions
8
Company profits

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.

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

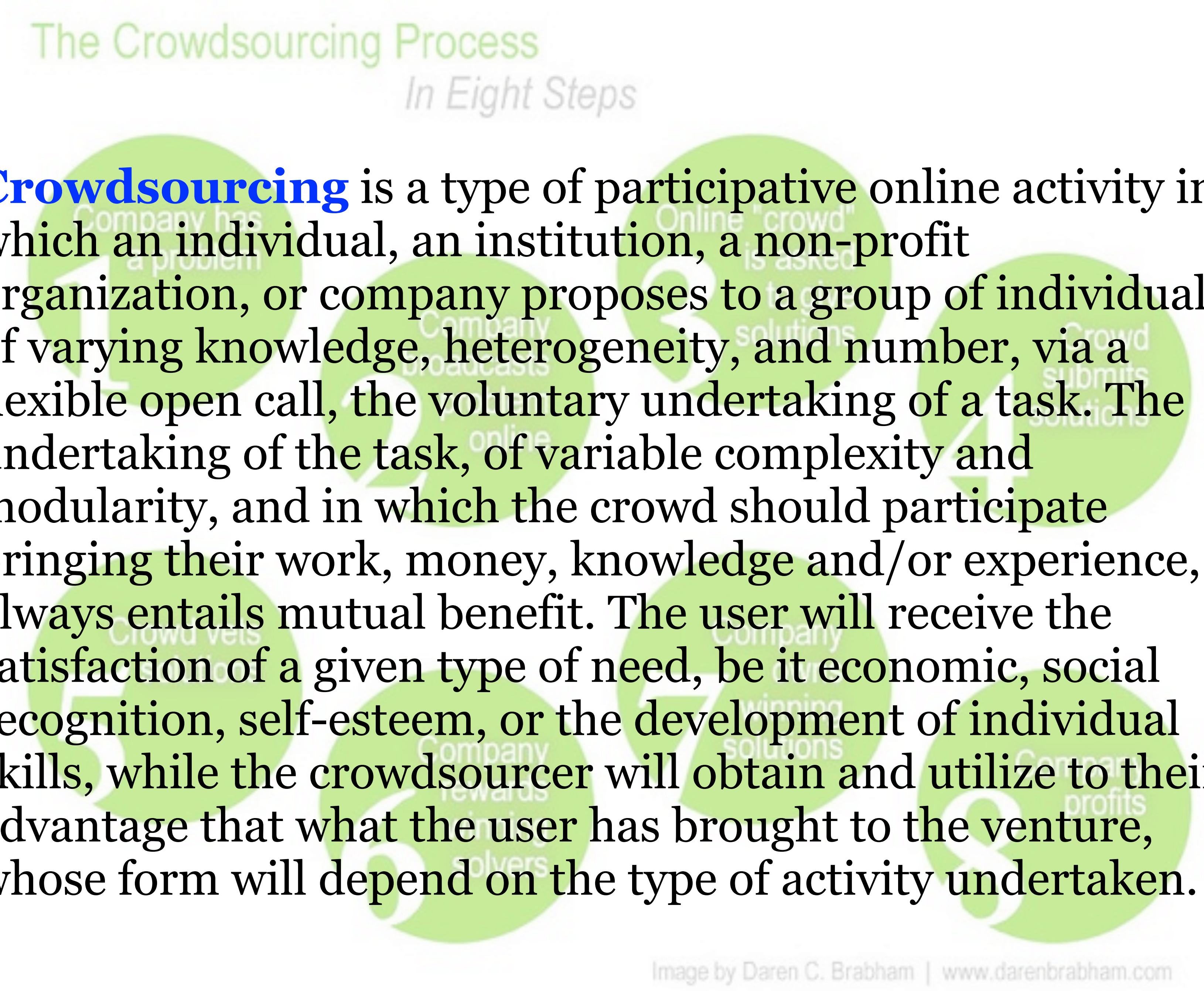
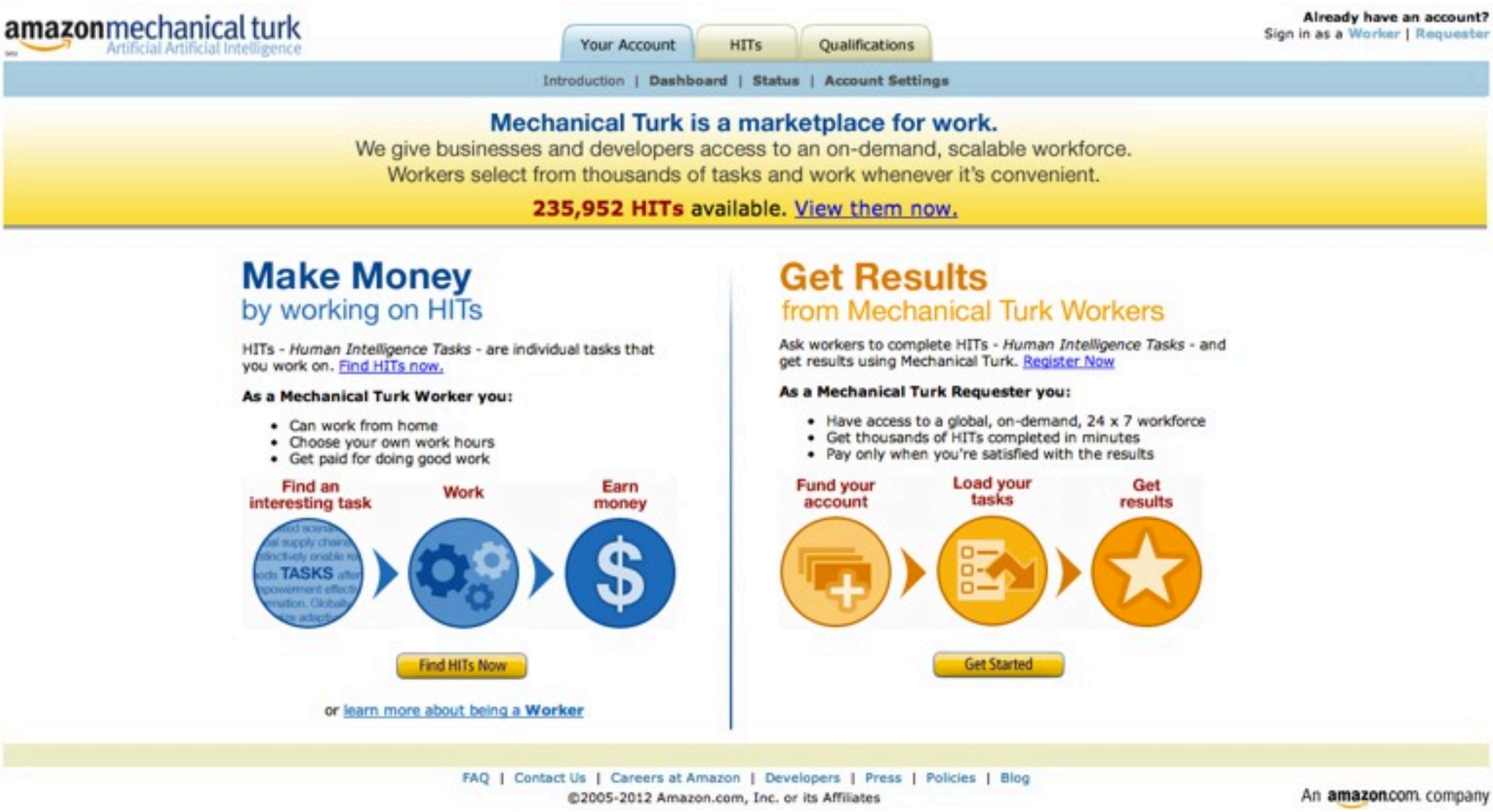

The Crowdsourcing Process
In Eight Steps
Company has a problem
Company broadcasts problem online
Online "crowd" is asked to submit solutions
Crowd submits solutions
Crowd vets solutions
Company rewards winning solvers
Company owns winning solutions
Company profits

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

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 XX(X). 2012. pp. 1-14.



amazonmechanical turk
Artificial Artificial Intelligence
Your Account
HITs
Qualifications
Already have an account?
Sign in as a Worker | Requester
Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

235,952 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

or learn more about being a **Worker**

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Register Now

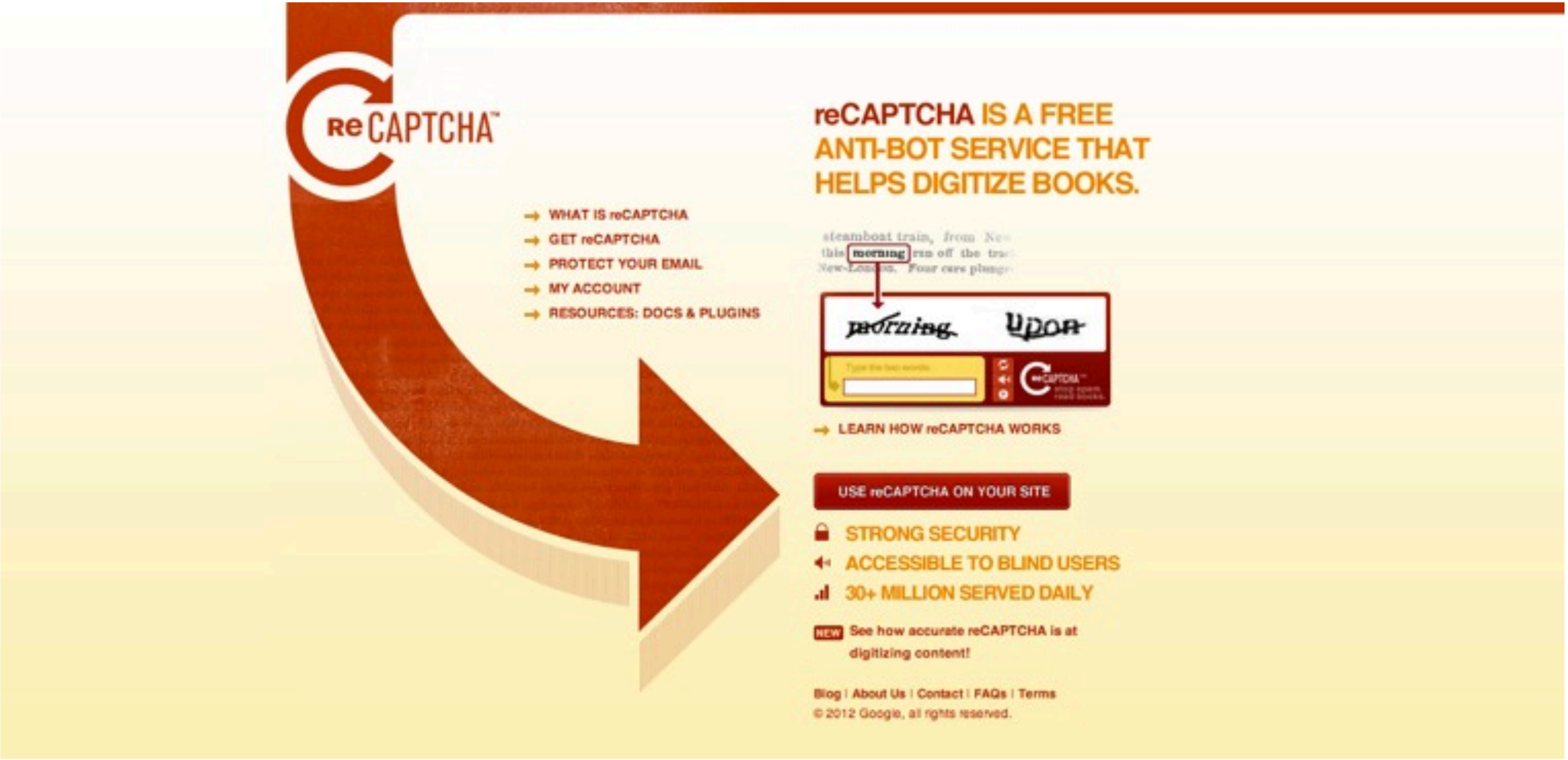
As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Fund your account
Load your tasks
Get results
Get Started
FAQ | Contact Us | Careers at Amazon | Developers | Press | Policies | Blog

©2005-2012 Amazon.com, Inc. or its Affiliates

Amazon Mechanical Turk was launched Nov 2005
Alexa global rank of Amazon Mechanical Turk (13-Jun-2012): 6,022



reCAPTCHA IS A FREE ANTI-BOT SERVICE THAT HELPS DIGITIZE BOOKS.

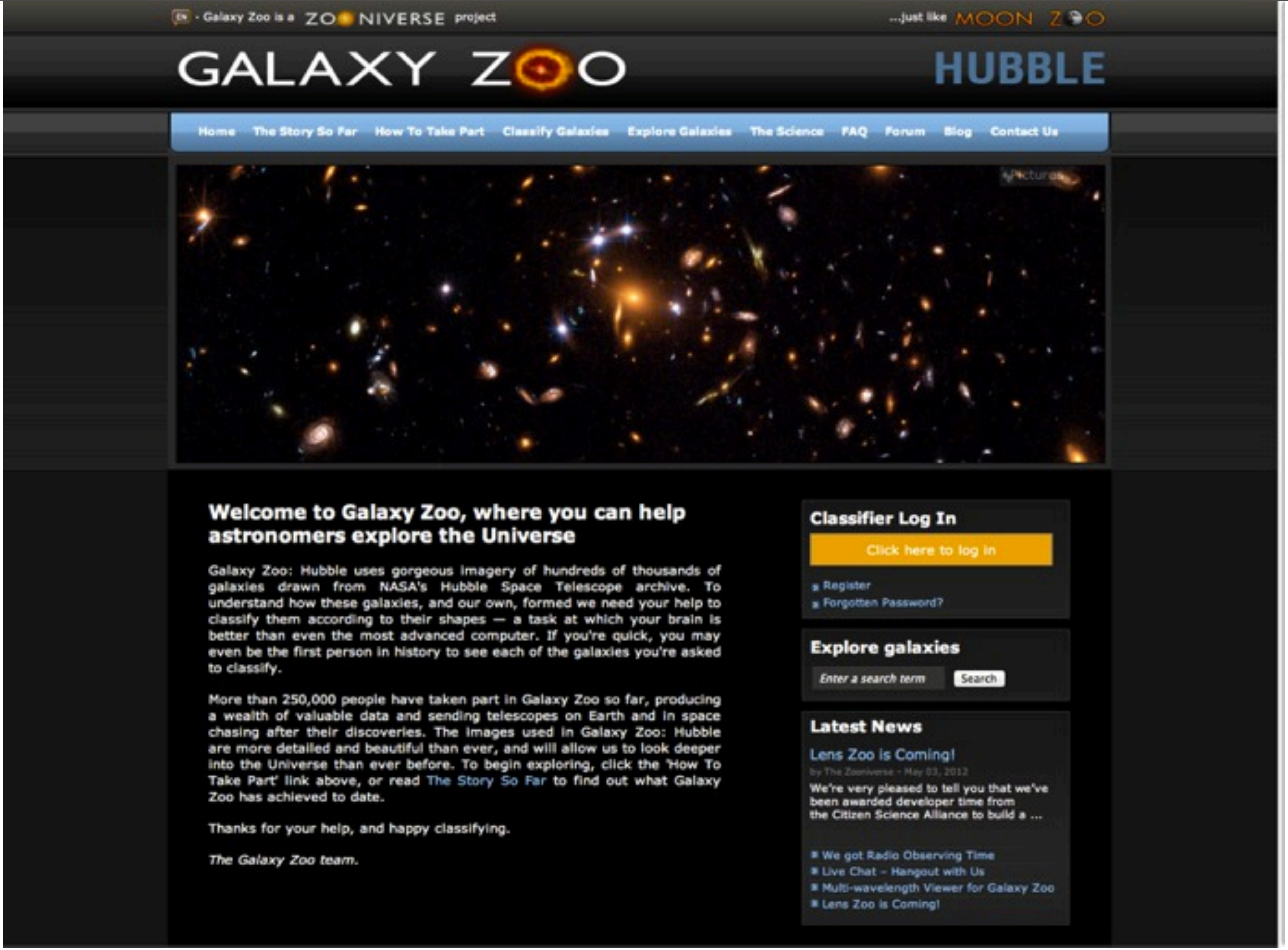
STRONG SECURITY

ACCESSIBLE TO BLIND USERS

30+ MILLION SERVED DAILY

NEW See how accurate reCAPTCHA is at digitizing content!

ReCAPTCHA™
WHAT IS reCAPTCHA
GET reCAPTCHA
PROTECT YOUR EMAIL
MY ACCOUNT
RESOURCES: DOCS & PLUGINS
steamboat train, from New
this morning run off the track
New-London. Four cars plunge
morning
upon
Type the two words:
reCAPTCHA™
stop spam. read books.
LEARN HOW reCAPTCHA WORKS
USE reCAPTCHA ON YOUR SITE
Blog | About Us | Contact | FAQs | Terms



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.

Latest News

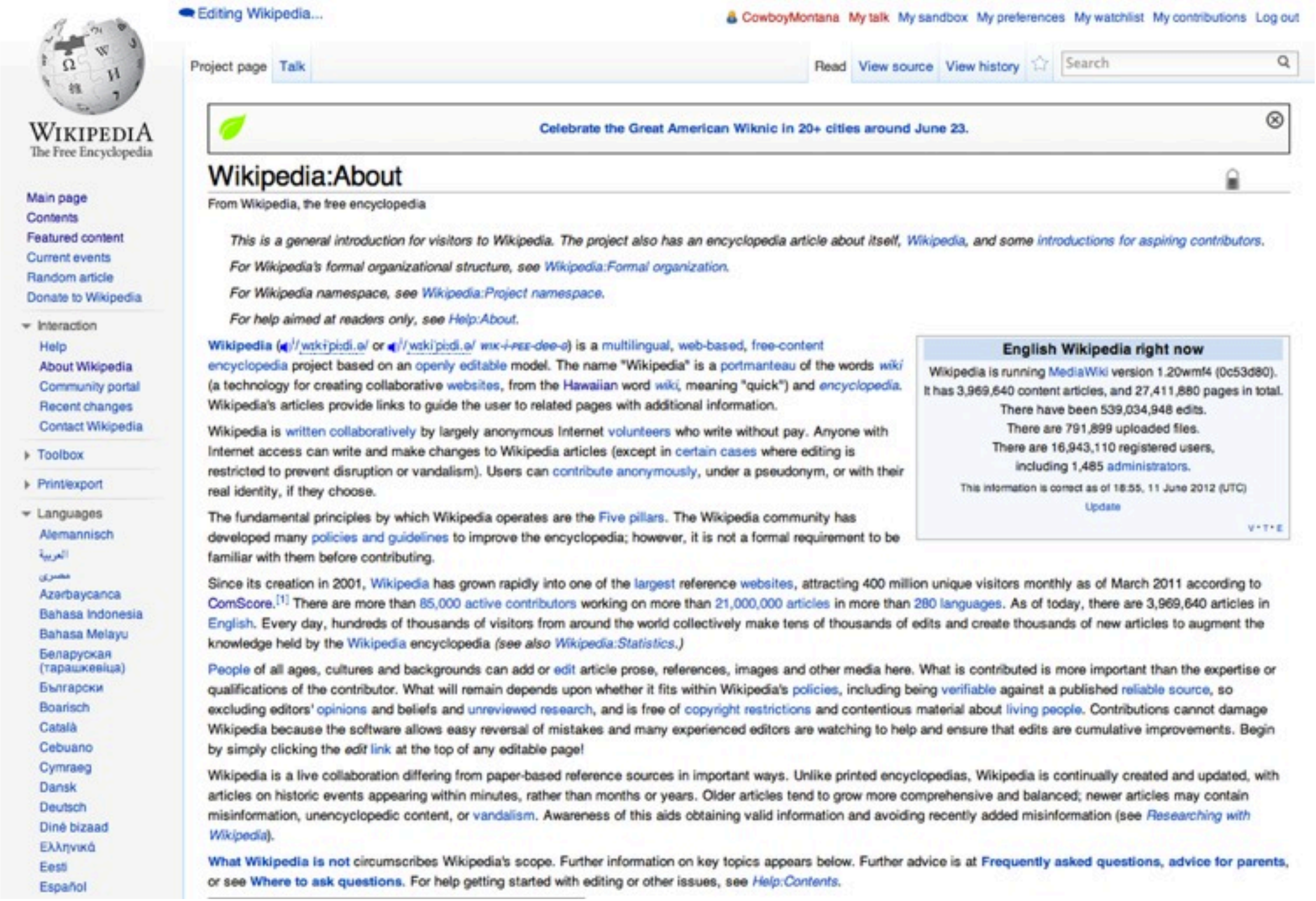
Lens Zoo is Coming!

by The Zooniverse - May 03, 2012

We're very pleased to tell you that we've been awarded developer time from the Citizen Science Alliance to build a ...

- We got Radio Observing Time
- Live Chat – Hangout with Us
- Multi-wavelength Viewer for Galaxy Zoo
- Lens Zoo is Coming!

Galaxy Zoo was 1st launched July 2007
Alexa global traffic rank of Galaxy Zoo (13-Jun-2012): 557,766



Celebrate the Great American Wiknic in 20+ cities around June 23.

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.

English Wikipedia right now

Wikipedia is running MediaWiki version 1.20wmf4 (0c53d80).
It has 3,969,640 content articles, and 27,411,880 pages in total.
There have been 539,034,948 edits.
There are 791,899 uploaded files.
There are 16,943,110 registered users,
including 1,485 administrators.
This information is correct as of 18:55, 11 June 2012 (UTC)
Update
V • T • E

Wikipedia (/ˌwɪkɪˈpiːdi.ə/ or /ˌwɪkiˈpiːdi.ə/ *WIK-i-PEE-dee-ə*) is a multilingual, web-based, free-content encyclopedia project based on an openly editable 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 write 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.)*

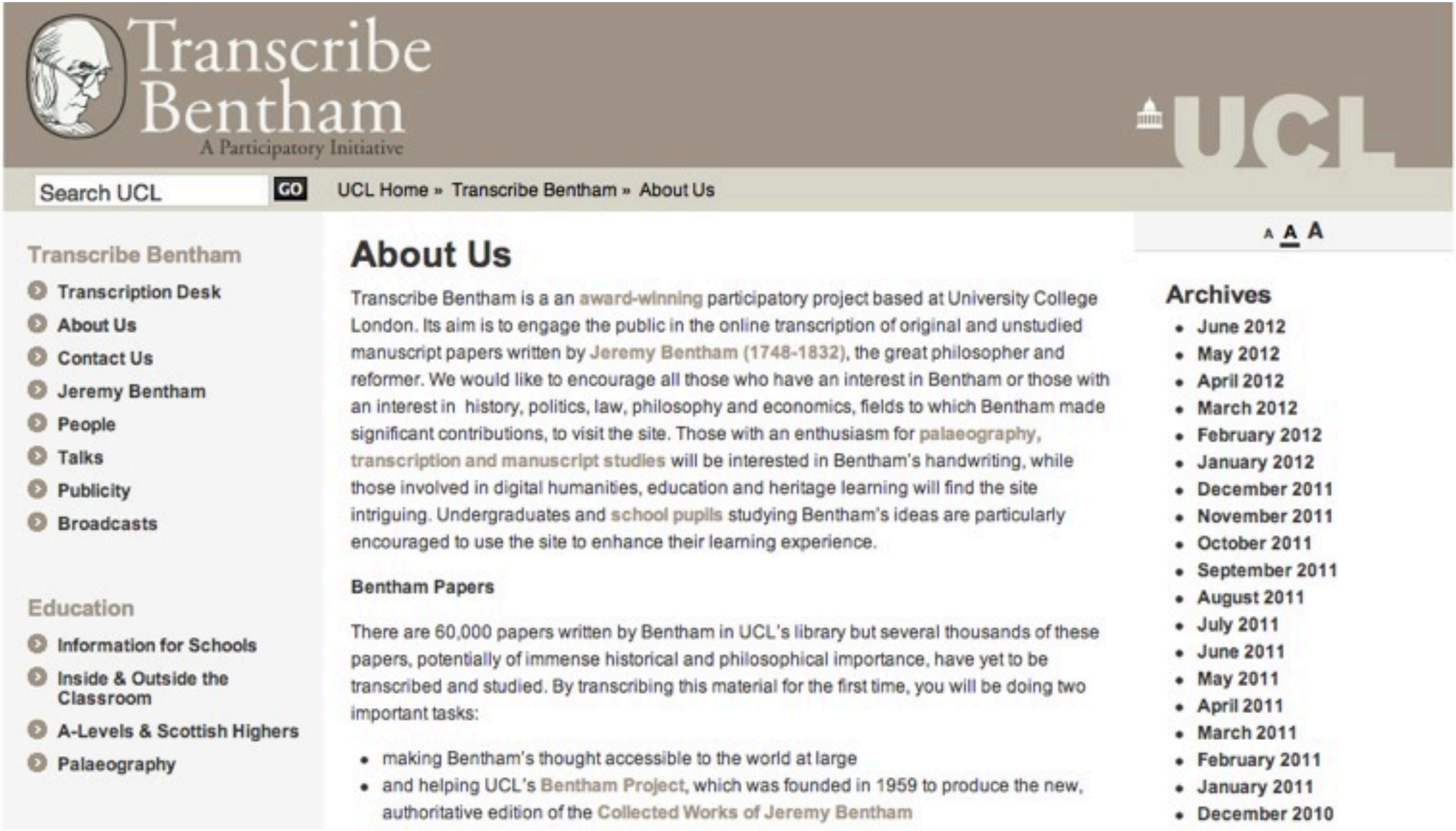
People of all ages, cultures and backgrounds can add or edit article prose, references, images and other media here. What is contributed is more important than the expertise or qualifications of the contributor. What will remain depends upon whether it fits within Wikipedia's policies, including being 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



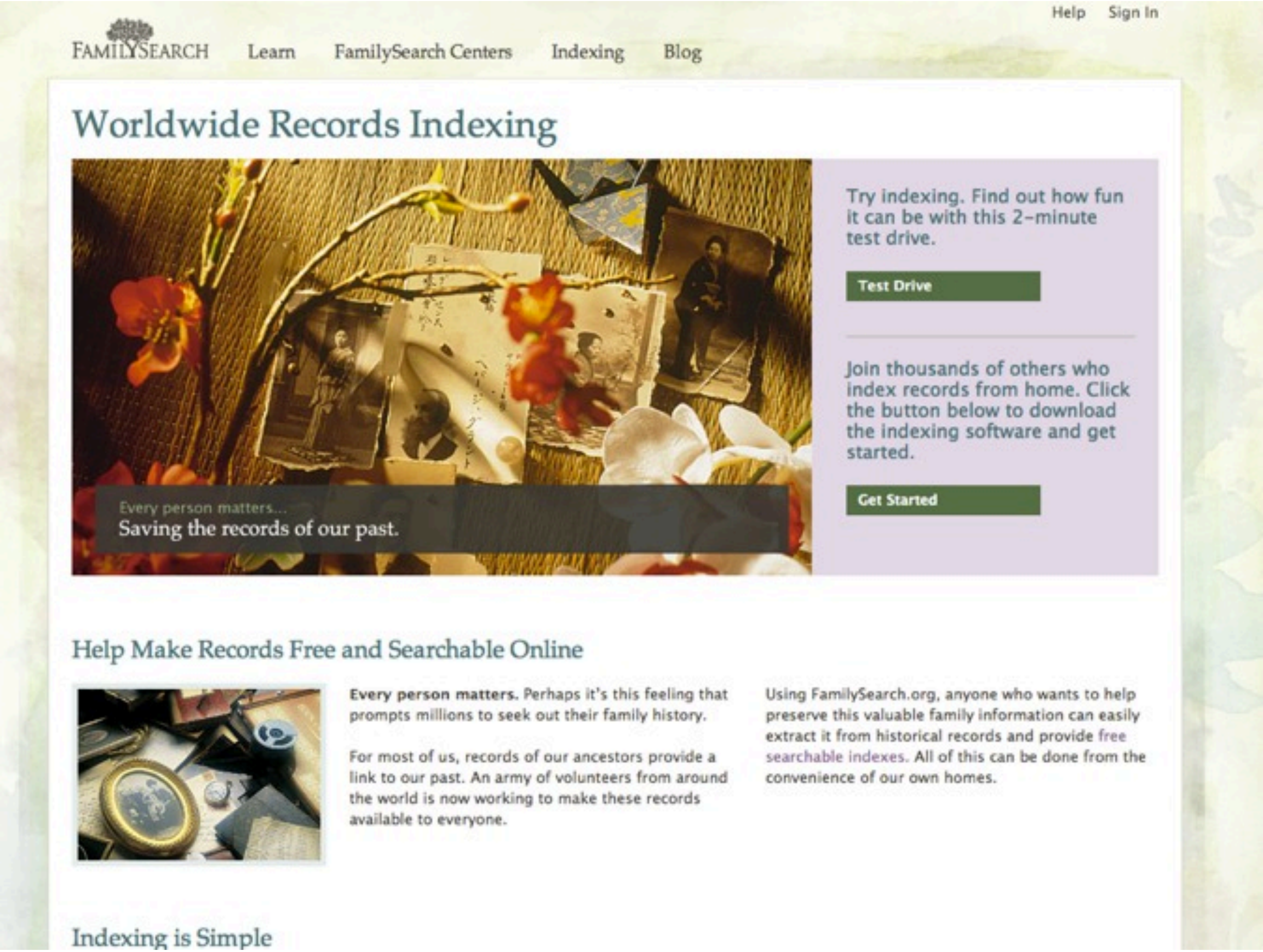
About Us

Transcribe Bentham is a 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
- and helping UCL's Bentham Project, which was founded in 1959 to produce the new, authoritative edition of the Collected Works of Jeremy Bentham



Worldwide Records Indexing

Every person matters...
Saving the records of our past.

Try indexing. Find out how fun it can be with this 2-minute test drive.

Join thousands of others who index records from home. Click the button below to download the indexing software and get started.

Help Make Records Free and Searchable Online

Every person matters. Perhaps it's this feeling that prompts millions to seek out their family history.

For most of us, records of our ancestors provide a link to our past. An army of volunteers from around the world is now working to make these records available to everyone.

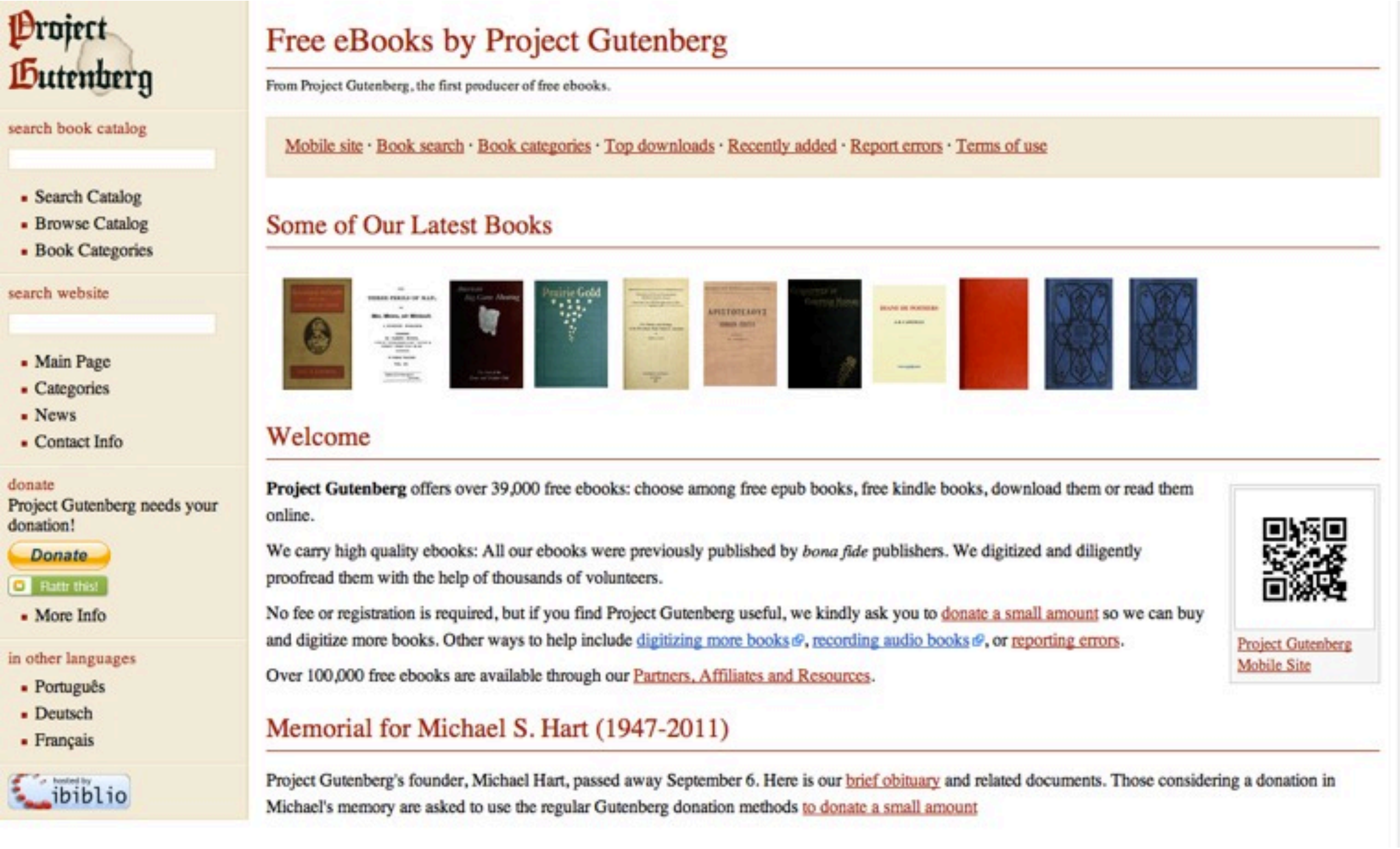
Using FamilySearch.org, anyone who wants to help preserve this valuable family information can easily extract it from historical records and provide free searchable indexes. All of this can be done from the convenience of our own homes.

Indexing is Simple

Family Search Indexing was 1st launched (beta) 2004
Alexa global traffic rank of FamilySearch (13-Jun-2012): 4,419

FamilySearch

- 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%



Free eBooks by Project Gutenberg

From Project Gutenberg, the first producer of free ebooks.

Some of Our Latest Books

Welcome

Project Gutenberg offers over 39,000 free ebooks: choose among free epub books, free kindle books, download them or read them online.

We carry high quality ebooks: All our ebooks were previously published by *bona fide* publishers. We digitized and diligently proofread them with the help of thousands of volunteers.

No fee or registration is required, but if you find Project Gutenberg useful, we kindly ask you to donate a small amount so we can buy and digitize more books. Other ways to help include digitizing more books, recording audio books, or reporting errors.

Over 100,000 free ebooks are available through our Partners, Affiliates and Resources.

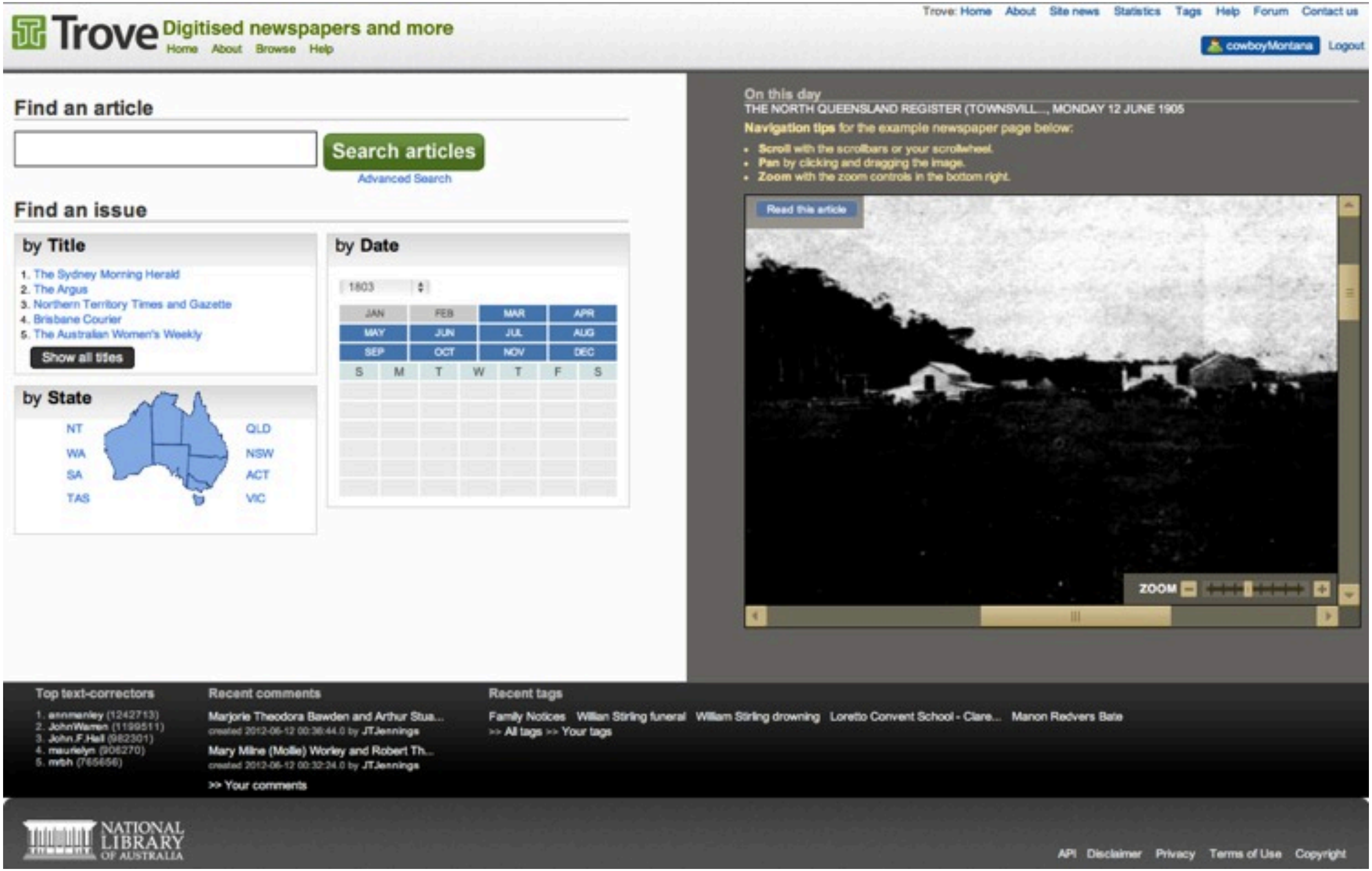
Memorial for Michael S. Hart (1947-2011)

Project Gutenberg's founder, Michael Hart, passed away September 6. Here is our brief obituary and related documents. Those considering a donation in Michael's memory are asked to use the regular Gutenberg donation methods to donate a small amount

Project Gutenberg was 1st launched Dec 1971
Alexa global traffic rank of Project Gutenberg (13-Jun-2012): 5,744

Project Gutenberg

- 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


Trove Digitised newspapers and more
Home About Browse Help
Trove: Home About Site news Statistics Tags Help Forum Contact us
cowboyMontana Logout
Search articles
Advanced Search
Show all titles
NT WA SA TAS QLD NSW ACT VIC
1803
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
S M T W T F S
Read this article
ZOOM
NATIONAL LIBRARY OF AUSTRALIA
API Disclaimer Privacy Terms of Use Copyright

Find an article

Find an issue

by Title

1. The Sydney Morning Herald
2. The Argus
3. Northern Territory Times and Gazette
4. Brisbane Courier
5. The Australian Women's Weekly

by State

by Date

On this day

THE NORTH QUEENSLAND REGISTER (TOWNSVILL..., MONDAY 12 JUNE 1905

Navigation tips for the example newspaper page below:

- **Scroll** with the scrollbars or your scrollwheel.
- **Pan** by clicking and dragging the image.
- **Zoom** with the zoom controls in the bottom right.

Top text-correctors

1. annmanley (1242713)
2. JohnWarren (1199511)
3. John.F.Hall (982301)
4. maurielyn (906270)
5. mrbh (765656)

Recent comments

Marjorie Theodora Bawden and Arthur Stua...
created 2012-06-12 00:36:44.0 by JTJennings

Mary Milne (Mollie) Worley and Robert Th...
created 2012-06-12 00:32:24.0 by JTJennings

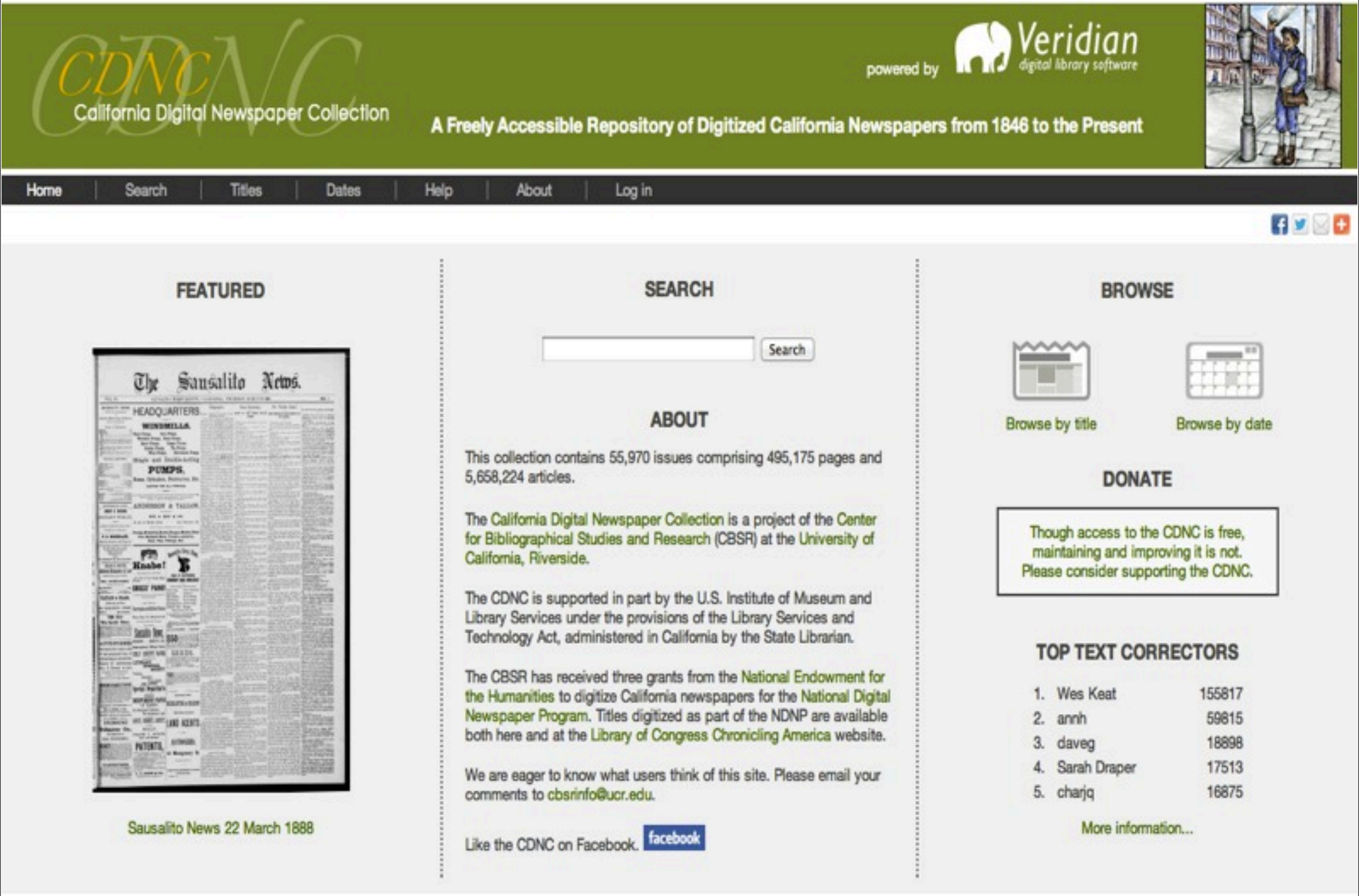
>> Your comments

Recent tags

Family Notices William Stirling funeral William Stirling drowning Loretto Convent School - Clare... Manon Redvers Bate
>> All tags >> Your tags

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)



CDNC

California Digital Newspaper Collection

powered by Veridian digital library software

A Freely Accessible Repository of Digitized California Newspapers from 1846 to the Present

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 cbsrinfo@ucr.edu.

Like the CDNC on Facebook. facebook

BROWSE

Browse by title

Browse by date

DONATE

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

TOP TEXT CORRECTORS

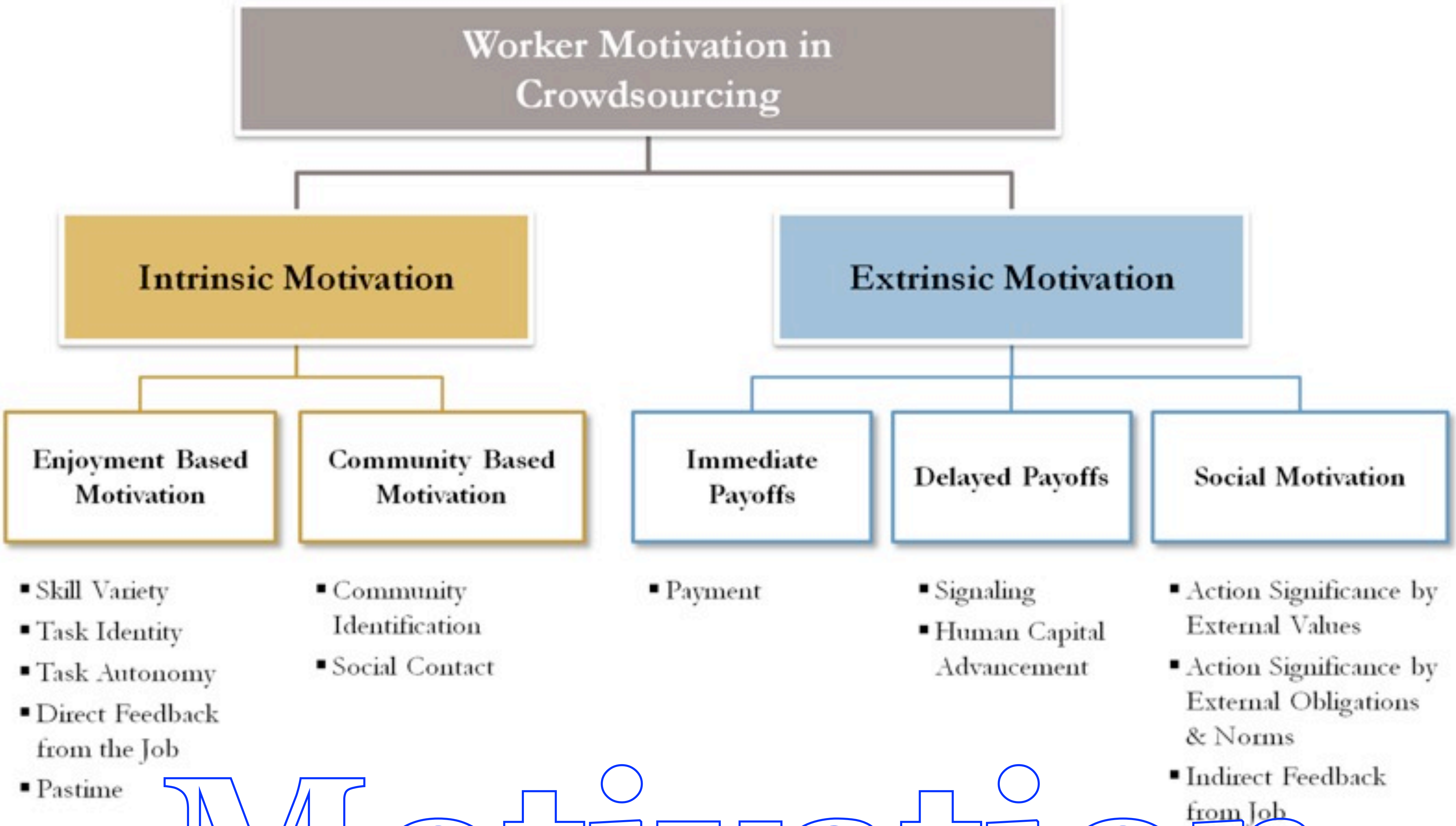
1.	Wes Keat	155817
2.	annh	59815
3.	daveg	18898
4.	Sarah Draper	17513
5.	charjq	16875

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

Motivation



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

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

Trove

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.”
National Library of Australia March 2009.

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”

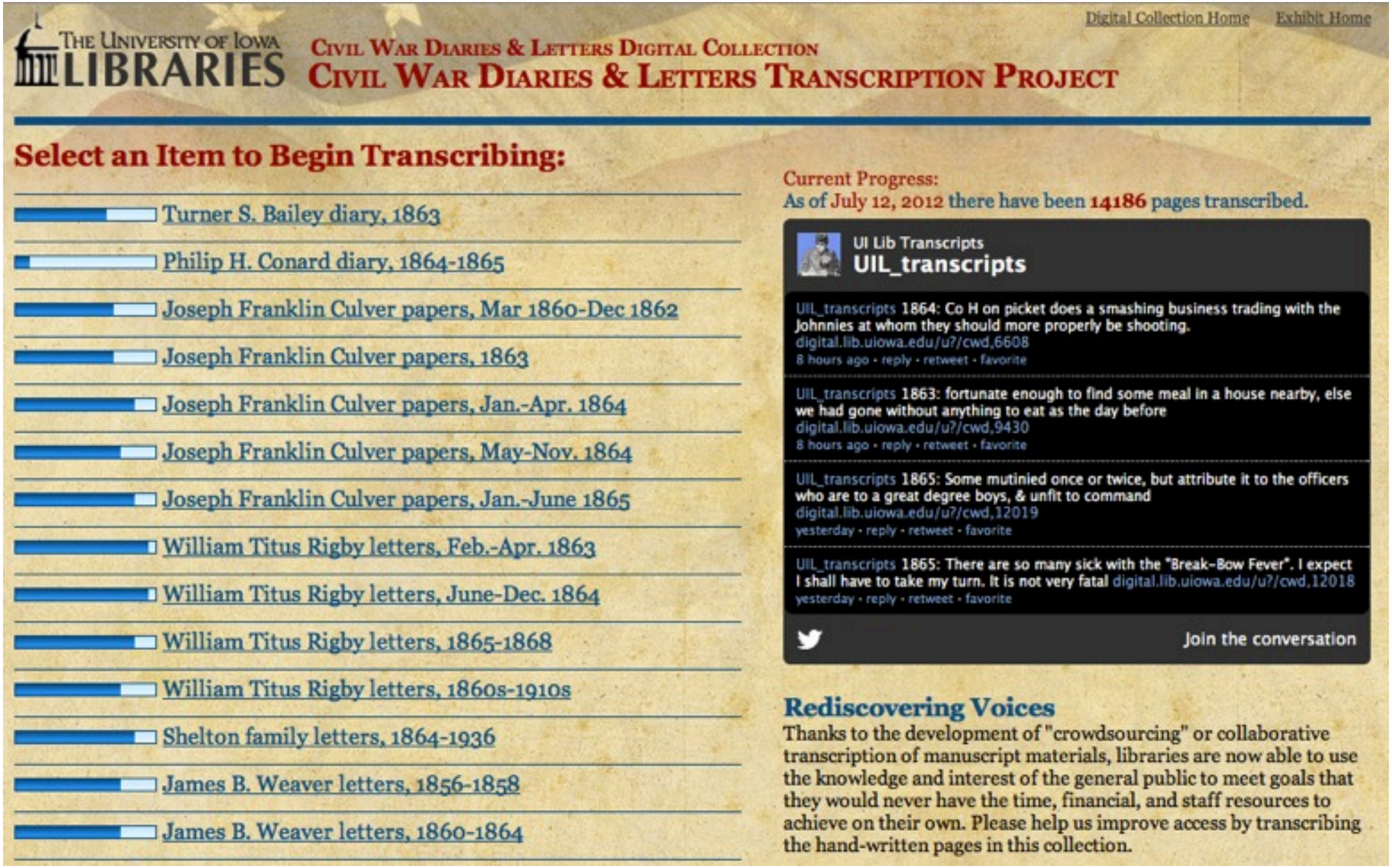
From Trevor Owen’s Crowdstorming blog
http://crowdstorming.wordpress.com/

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



Digital Collection Home Exhibit Home

The University of Iowa LIBRARIES

Civil War Diaries & Letters Digital Collection

Civil War Diaries & Letters Transcription Project

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.

UI Lib Transcripts
UIL_transcripts

UIL_transcripts 1864: Co H on picket does a smashing business trading with the Johnnies at whom they should more properly be shooting.
digital.lib.uiowa.edu/u?/cwd,6608
8 hours ago · reply · retweet · favorite

UIL_transcripts 1863: fortunate enough to find some meal in a house nearby, else we had gone without anything to eat as the day before
digital.lib.uiowa.edu/u?/cwd,9430
8 hours ago · reply · retweet · favorite

UIL_transcripts 1865: Some mutinied once or twice, but attribute it to the officers who are to a great degree boys, & unfit to command
digital.lib.uiowa.edu/u?/cwd,12019
yesterday · reply · retweet · favorite

UIL_transcripts 1865: There are so many sick with the "Break-Bow Fever". I expect I shall have to take my turn. It is not very fatal digital.lib.uiowa.edu/u?/cwd,12018
yesterday · reply · retweet · favorite

Join the conversation

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 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.

THE UNIVERSITY OF IOWA

Website traffic

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

	11-Jun-2011 / 12-Jul-2011	11-Jun-2012 / 12-Jul-2012	change
visits	16,934	20,758	+22.6%
unique visitors	11,030	12,951	+17.4%
visit duration	9m 24s	11m 6s	+18.1%
bounce rate	51.3%	44.7%	-12.9%

Crowdsourcing benefits

Public domain photo courtesy of US Navy

Economics

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 x 35 characters per line x 1/1000 x $0.50 = $6,901 $$
- Trove: 69,918,892 lines x 35 characters per line x 1/1000 x $0.50 = $1,223,581 $$$$$

Text accuracy

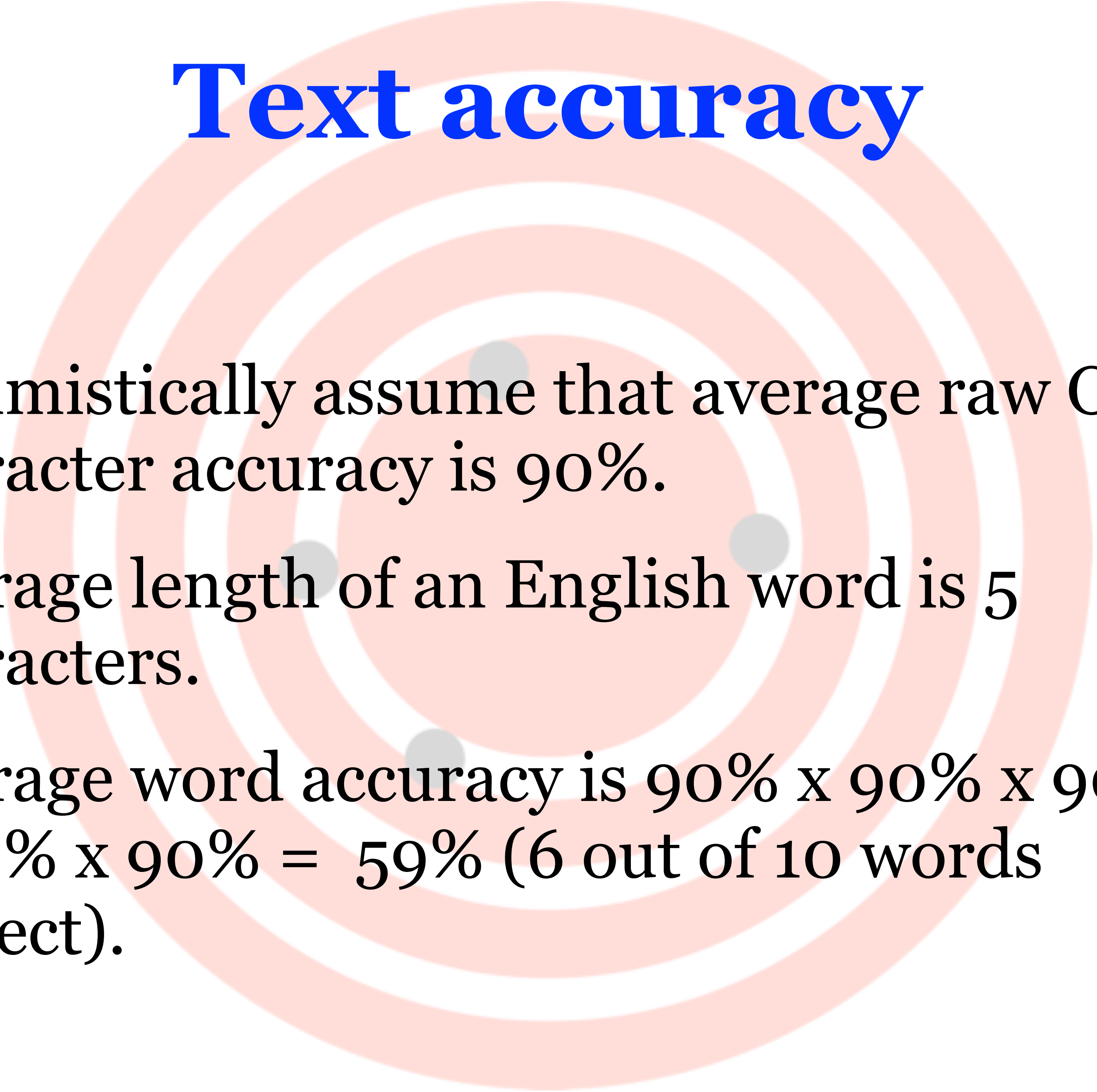
- Edwin Kiljin (Koninklijke Bibliotheek the Netherlands) reports raw OCR character accuracies of 68% for early 20th century newspapers

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

Edwin Kiljin. “The current state-of-art in newspaper digitization.” D-Lib Magazine. January/February 2008.

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.

Text accuracy



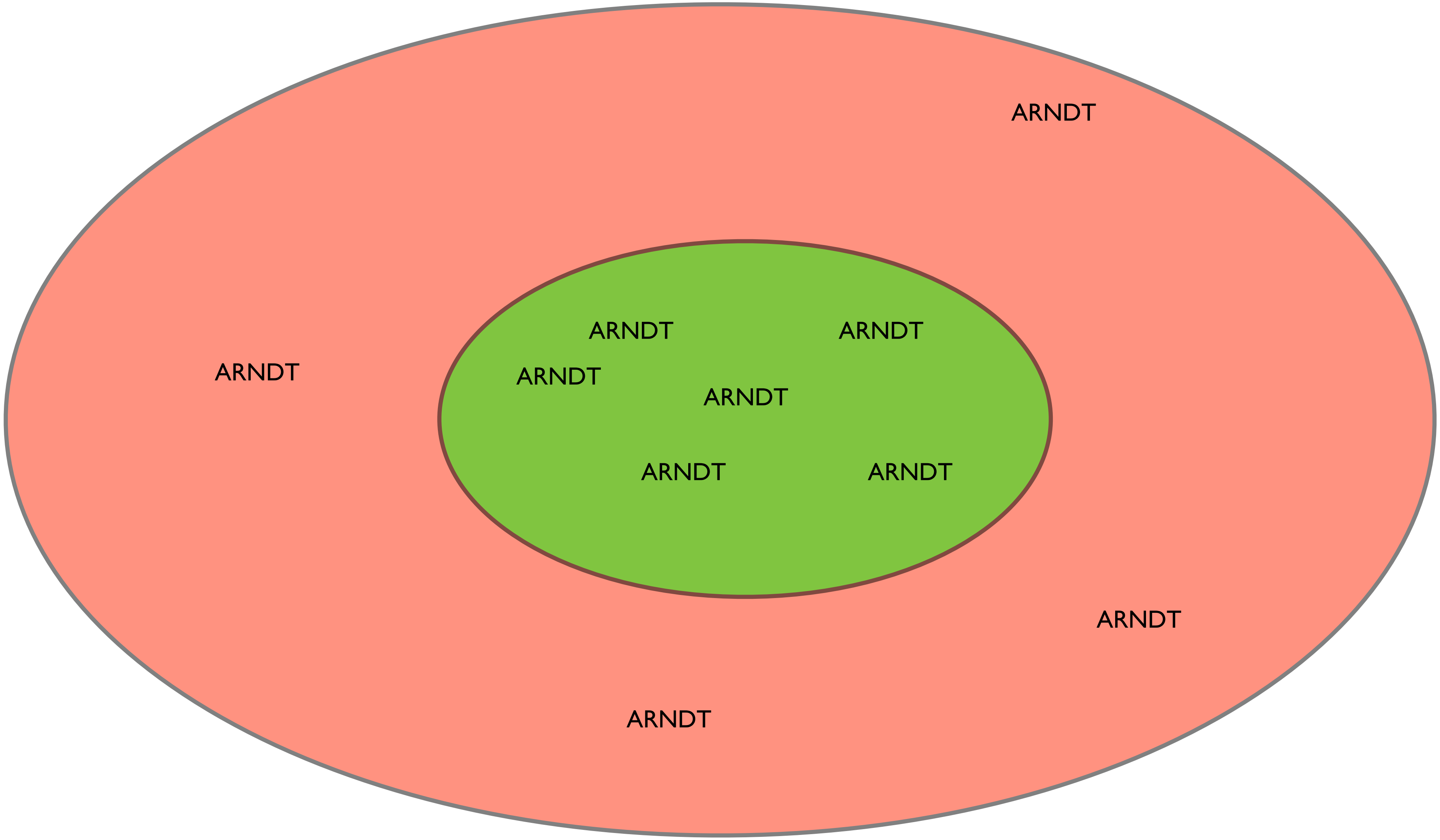
Optimistically assume that average raw OCR character accuracy is 90%.

Average length of an English word is 5 characters.

Average word accuracy is 90% x 90% x 90% x 90% x 90% = 59% (6 out of 10 words correct).

Public domain graphic courtesy of Wikimedia Commons.

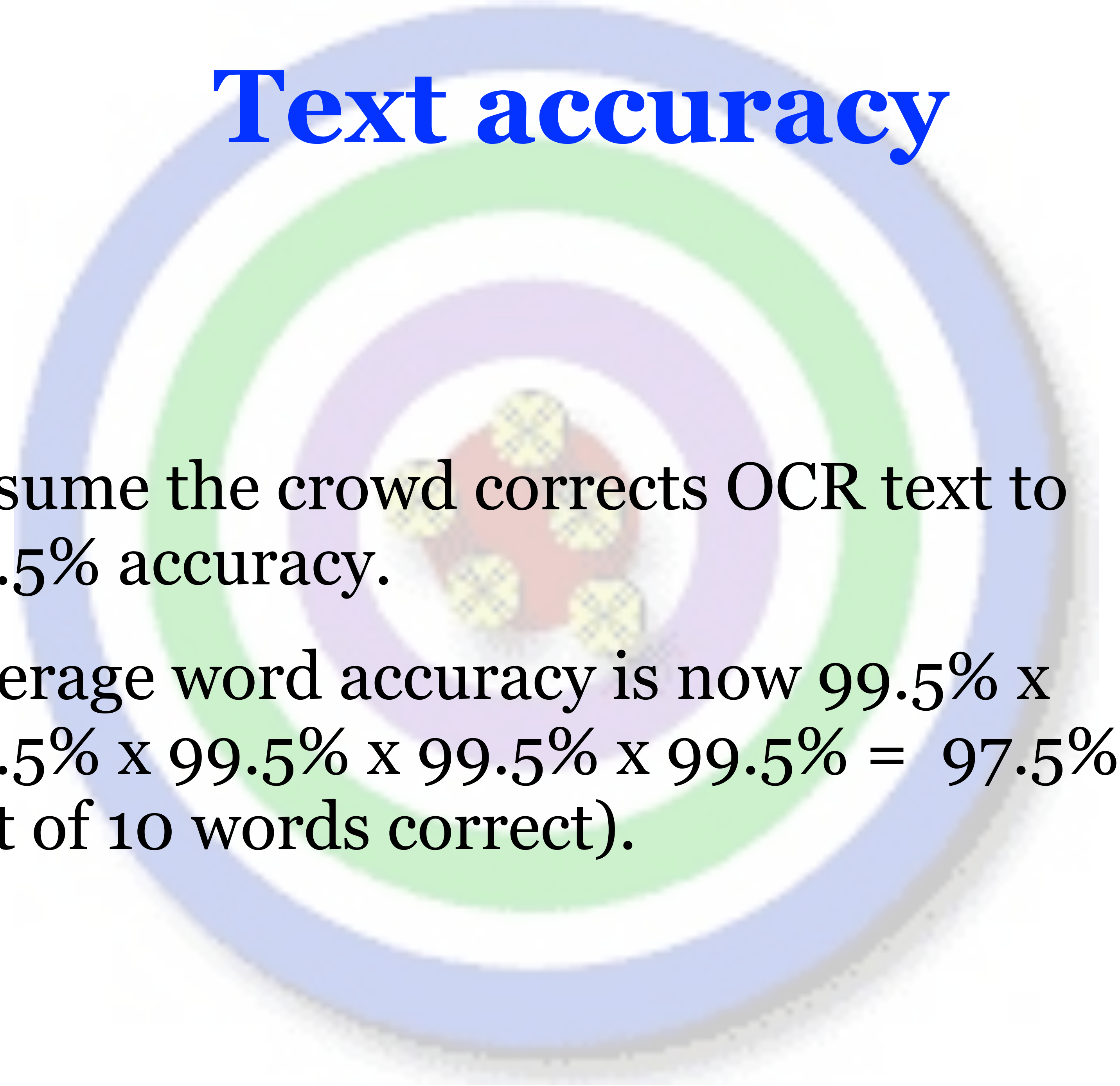
Search recall
no text correction



instances of “ARNDT” found

instances of “ARNDT” not found

Text accuracy

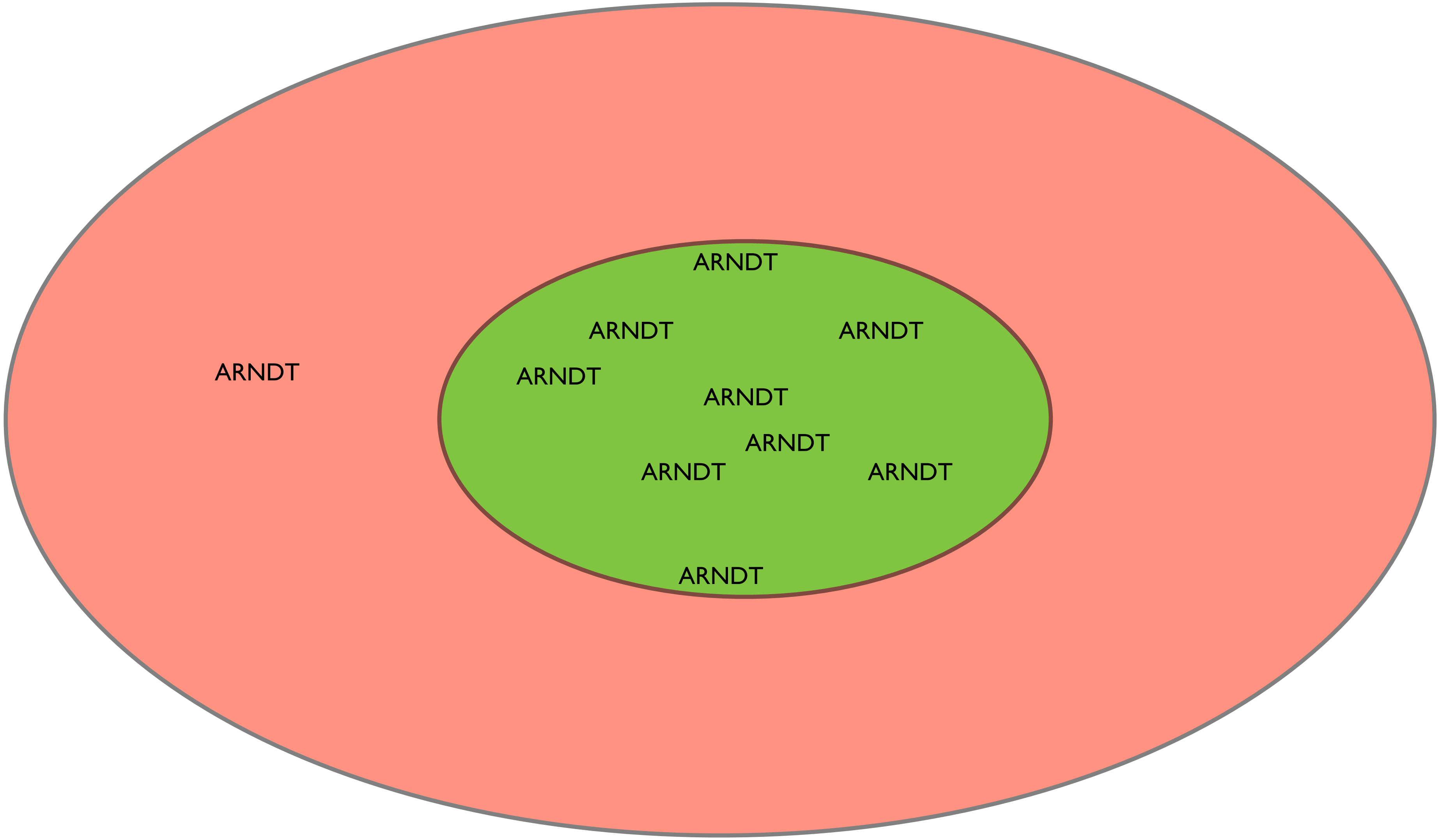


Assume the crowd corrects OCR text to 99.5% accuracy.

Average word accuracy is now 99.5% x 99.5% x 99.5% x 99.5% x 99.5% = 97.5% (6 out of 10 words correct).

Public domain graphic courtesy of Wikimedia Commons.

Search recall with text correction



instances of "ARNDT" found

instances of "ARNDT" not found

?

Frederick Zarndt

frederick@frederickzarndt.com

sponsored by

CCS / Digital Divide Data / DL Consulting

Photo held by John Oxley Library, State Library of Queensland. Original from Courier-mail, Brisbane, Queensland, Australia.