Spring 2014

Internet Reviews: Crowdsourcing in Libraries and Archives

Jennifer A. Bartlett
University of Kentucky, jen.bartlett@uky.edu

INTERNET REVIEWS: CROWDSOURCING IN LIBRARIES AND ARCHIVES

BY JENNIFER A. BARTLETT

HEAD OF W.T. YOUNG LIBRARY REFERENCE SERVICES
UNIVERSITY OF KENTUCKY LIBRARIES

Imagine a vast army of workers ready to help you comb through your online print, audio and video collections, transcribing and correcting text, identifying the subject of photographs and providing subject tagging to item records. Also, you don't have to pay them a thing.

Enter crowdsourcing,[1] a distributed work process in which tasks are outsourced to a large group of people working at different locations and at their own speed. Rather than belonging to a specified group of employees or contractors, people who work on crowdsourced projects are either volunteers or part-time freelancers who generally work online and from home.

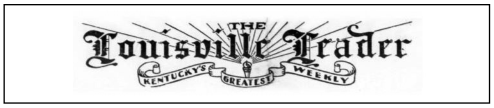
An often-cited example of crowdsourcing is **Amazon's Mechanical Turk** (https://www.mturk.com). Mechanical Turk is an example of "microwork," or tiny tasks that take little time and pay very low amounts of money. For example, a worker is asked to transcribe text from business cards for $0.02 per card, tag images with subject terms for $0.04 per five-image set, or even to write a brief industry trend report for $22.50. As the complexity of the task increases, so does the amount paid.

Interestingly, microworked crowdsourcing is not always done consciously. Many websites utilize CAPTCHA ("Completely Automated Public Turing test to tell Computers and Humans Apart"), a program that helps determine whether a website user is a human or a computer by asking the user to type distorted text. A related service, reCAPTCHA, uses this technology to protect sites against spam and malicious attacks while simultaneously helping to digitize archival books and newspapers. Drawing from digitized content including editions of the *New York Times* and books from Google Books (Google owns the service), reCAPTCHA requires users to type two words, one known and one unreadable by OCR. If the user types the known word correctly, the system assumes the answer for the second to be correct as well. Repeating the process helps reCAPTCHA develop an accurate group consensus. According to a 2008 article in *Science* magazine,[2] the success rate for word identification through reCAPTCHA is 96.1%.

Crowdsourcing is enjoying increasing popularity in a number of industries, including architecture and urban planning, technology, advertising, graphic design, and more.[3] A proliferation of archival collections being made available online translates into an opportunity for libraries and other cultural institutions to tap into a potential pool of content editors and annotators. Optical character recognition (OCR) scanning methods, while efficient, are often riddled with errors due to the poor quality and condition of the original source images. Identification of handwritten text can also be problematic. A growing number of libraries are getting assistance from their online patrons in transcribing text, identifying images and other content, and tagging elements in digitized documents through crowdsourcing platforms. Using crowdsourcing, large quantities of digitized content quickly becomes searchable for researchers and general users. The following projects are just a few examples of successful crowdsourcing in libraries and archives.

NEWSPAPERS

University of Louisville: The Louisville Leader
http://digital.library.louisville.edu/cdm/landingpage/collection/leader/
The Louisville Leader, "Kentucky's Greatest Weekly," was an African-American community newspaper published from 1917 to 1950,



and covered a wide range of local educational, social, religious and other topics in addition to national and international news. The iTranscribe site includes featured articles scanned from microfilm, and is based on the open-source publishing platform Omeka (http://omeka.org/), using the Scripto plug-in for transcription (http://scripto.org/). For more information, see project coordinator Rachel Howard's "Genealogy Gems" column in the Spring 2014 issue of *Kentucky Libraries*.

California Digital Newspaper Collection
http://cdnc.ucr.edu/about_us.html

Covering historical California newspapers from 1846 to 1922, the CDNC contains over 40,000 pages of content including articles from the *Californian*, the first California newspaper, and the *Daily Alta California*, the first daily. CDNC also provides access to several contemporary state newspapers as part of a pilot project, for example, the *LA Downtown News* and the *Santa Cruz Sentinel.*

The site, based on Digital Library Consulting's Veridian platform (http://veridiansoftware.com/crowdsourcing/), does not widely advertise its crowdsourcing functionality, but does offer a list of "top text correctors" on its front page. Transcribers can register for an account on the site, and immediately begin transcribing and correcting text. As of this writing, there are over 2,000 registered users, with nearly 2.4 million lines of text corrected.

National Library of Australia: Trove Australian Newspapers
http://trove.nla.gov.au

The National Library of Australia's Trove system is home to millions of items including pictures, music and sound clips, maps, theses, archived websites, and more. Trove provides access to over 12 million pages from more than 650 pre-1955 Australian newspapers. The crowdsourcing program associated with the newspaper digitization program is very well organized, offering an extensive step-by-step FAQ, up-to-date correction and tagging statistics, and even a Trove forum in which contributors can talk about news, uses for Trove content, and problems.

HISTORICAL COLLECTIONS

University of Alabama: "Tag It – A Historical Photograph Tagging Project"
tagit.lib.ua.edu/

University of Alabama Libraries
Name. tag it
Description. A historical photograph tagging project.

Librarians at the University of Alabama encourage their users to add relevant information to their thousands of digitized images through the "Tag It" project. In the hopes of drawing on users' local historical knowledge, "Tag It" provides an interface through which users can annotate the sometimes incomplete descriptions of some archival images with geographic locations, personal names, key words, and other descriptors. Tags, which are not limited to a controlled vocabulary or thesaurus, are also added to Acumen, the Libraries' digital repository, for searching by other users. The Libraries also invites transcription of many of its hand-written collections at http://transcribe.lib.ua.edu/.

The University of Iowa: DIY History
http://diyhistory.lib.uiowa.edu/

In the spring of 2011, the University of Iowa Libraries opened its Civil War Diaries and Letters Transcription Project to the public to commemorate the Civil War sesquicentennial. One year and over 15,000 transcribed pages later, the Civil War project was nearly completed and the project was expanded to include other archival materials. In October 2012, DIY History was launched. According to project developers, the goal of DIY History is to "make historic artifacts more accessible – both by enhancing catalog records for greater ease in searching and browsing, and by engaging the public to interact with the materials in new ways." Digital content on the site, which uses Omeka with the Scripto plug-in, comes from the Iowa Digital Library and features selected documents from the University's Special Collections, Archives, and the Iowa Women's Archives.

BOOK AND RECORD TRANSCRIPTION

Project Gutenberg: Distributed Proofreaders
http://www.pgdp.net/c/

One of the best-known and earliest crowdsourcing projects is Project Gutenberg, founded by Michael Hart in 1971 and still the largest single collection of free electronic books. It continues its original mission of "encouraging the creation and distribution of ebooks" through a network of volunteers. Distributed Proofreaders, a nonprofit organization run entirely by volunteers, was founded in 2000 to support Project Gutenberg, and currently is the major source of PG titles. "Distributed proofreaders" find, scan, and mark up books in the public domain, page by page. Rather than one volunteer being responsible for the production of an entire work, the distributed proofreading process divides each work into individual pages which can be proofread by several volunteers before the e-book is completed. The Distributed Proofreaders organization claims to have 90,000 volunteers who have made available 16,000 texts through Project Gutenberg over the past ten years.

The Church of Jesus Christ of Latter Day Saints: FamilySearchIndexing
https://familysearch.org/indexing/

Familysearch, a nonprofit genealogical organization, offers several family research services to content providers including image capture, digital conversion and online indexing. Through FamilySearchIndexing, volunteers can work on transcribing records in over 100 projects worldwide, including 19th century Catholic Church records in Saskatchewan, census records in Ghana, and passenger lists from Boston, Massachusetts. Common materials to transcribe include documents such as birth, death, marriage and census records. The resulting millions of documents are included in the free online FamilySearch database at https://familysearch.org/search.

New York Public Library: What's on the Menu?
http://menus.nypl.org/

What were New Yorkers eating in 1940? A good place to look is the NYPL's extensive restaurant menu collection, comprised of over 45,000 items from the mid-1800's to the present, which also includes other items related to food lore. Historians, chefs and food enthusiasts value this well-presented collection for its information about not only names of restaurants and menus, but details about specific dishes, prices, and other historical information. Volunteers comb through menus for information beyond standard descriptive cataloging data, such as the name of the restaurant, geographical location, and the items and prices on the menus themselves.

"What's on the Menu?" is a project of NYPL Labs, a digital research experimental and technology center (http://www.nypl.org/collections/labs). Other crowdsourced projects from the Labs include "Building Inspector" (http://buildinginspector.nypl.org/), a web app that utilizes volunteers to check information from 19th century New York City insurance maps, and "Direct Me NYC: 1940" (http://directme.nypl.org/), in which volunteers created targeted searches of 1940 census data against addresses found in NYC telephone directories.

The possibilities of crowdsourced projects for libraries and archives are indeed intriguing, but the model is not without potential problems. The term "crowdsourcing" itself can connote exploitation of volunteers, and raises the problem of lack of quality control. However, in practice, the "crowds" in crowdsourcing are generally a group of dedicated, interested people who are committed to the idea of cultural stewardship. With appropriate training and oversight, crowdsourcing projects can be an excellent way to encourage patron collaboration and cooperation, foster a sense of public ownership, complete projects that the library might not have the resources to accomplish otherwise, and add value to our collections.

Jennifer Bartlett
jen.bartlett@uky.edu

FOOTNOTES

[1] The term "crowdsourcing," a combination of "outsourcing" and "crowd," was coined in a 2006 *Wired* magazine article, "The Rise of Crowdsourcing," by writer Jeff Howe (http://archive.wired.com/wired/archive/14.06/crowds.html).

[2] Von Ahn, Luis, et al. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321: 1465-8. Available at http://www.google.com/recaptcha/reCAPTCHA_Science.pdf

[3] For a representative list of more projects, see Wikipedia's "List of Crowdsourcing Projects" at http://en.wikipedia.org/wiki/List_of_crowdsourcing_projects. Of course, Wikipedia itself is an example of crowdsourced knowledge accumulation.