Experimenting with Strategies for Crowdsourcing Manuscript Transcription

Nicole Saylor, Head, Digital Library Services, University of Iowa Libraries
Jen Wolfe, Metadata Librarian, University of Iowa Libraries

Introduction

Crowdsourcing—soliciting the public's help to perform a task—is a creative way to garner a workforce to help transcribe, annotate, measure, and rectify archival materials. Social media tools are making this possible on the necessary scale, something prohibitively expensive by conventional means. This public engagement not only results in free labor for libraries, but it allows users to interact with library materials in a whole new way. Citizen contributors can follow the stories revealed by historic documents. Some become invested in those stories or motivated by furthering the mission of research by enhancing access to important historic documents.

While the crowdsourced contributions are free, the projects are by no means without cost, especially in regards to staff time. At the University of Iowa (UI), the Digital Library department reluctantly turned down an initial request from curators to develop a crowdsourcing initiative for transcribing Civil War diaries, citing a lack of sufficient programming expertise. That decision was revisited, however, thanks to creative thinking on the part of key staff members, and the UI's Civil War Diaries and Letters Transcription Project (http://digital.lib.uiowa.edu/cwd/) was launched in the spring of 2011. Six months later the effort is, by many measures, a certified success. Early response was so enthusiastic it crashed the Digital Library server, and today a devoted stable of transcribers continues to contribute to the project.

Wide-Ranging Options

Inspired by international and non-library efforts, libraries in the US are experimenting with a range of crowdsourcing tools. At one end of the spectrum are projects that use free and cloud-based solutions, such as the North Carolina State Library Family History project,[1] which seeks user-generated transcriptions through Flickr. Oral Roberts University's Written Rummage project[2] to transcribe a Fredrick Douglass diary uses a free Amazon cloud service, Mechanical Turk, to manage the transcription workflow.[3] Other more well-heeled projects have resulted in efforts such as What's on the Menu?,[4] a New York Public Library project to transcribe historic restaurant menus. As of the end of November 2011, there have been 645,517 dishes transcribed from 10,960 menus.

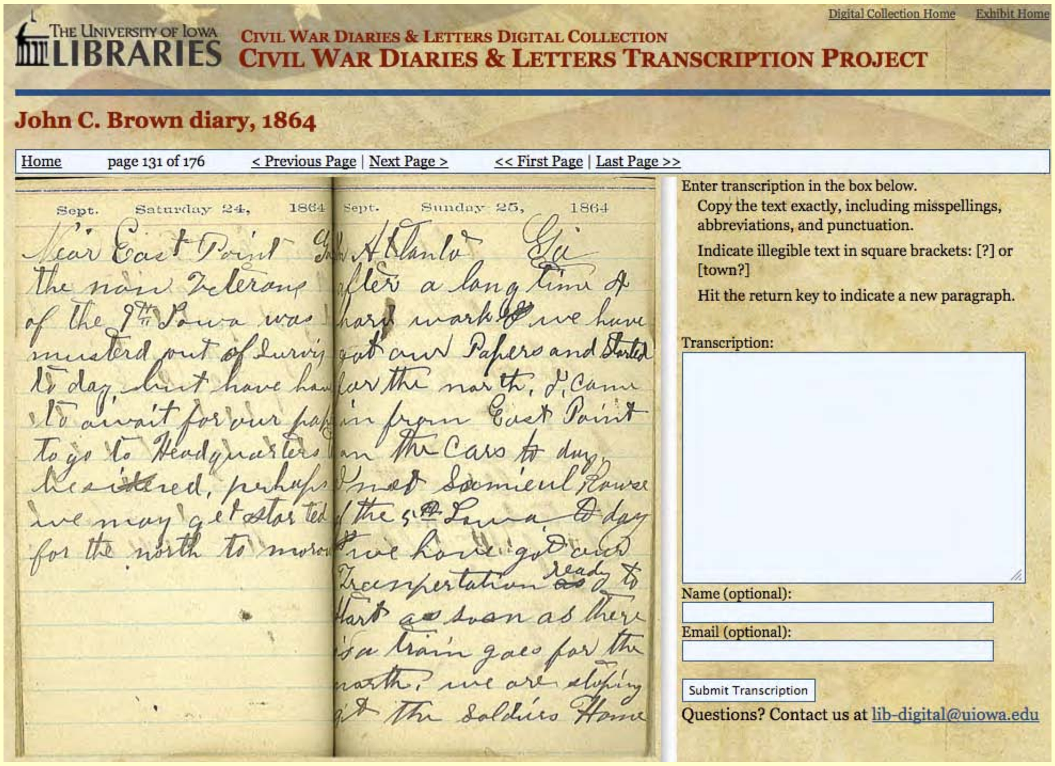
Some projects rely on specialized crowdsourcing software. Among the first to enter this arena was software engineer Ben Brumfield, who built the web-based tool From the Page[5] for transcribing, indexing, and annotating handwritten material. At the time Iowa was starting its project, this was the only open-source solution around. Since then, with the help of grants from the National Endowment for the Humanities Office of Digital Humanities, the Roy Rosenzweig Center for History and New Media has developed an open-source tool Scripto and applied it to transcribe 45,000 papers of the War Department.[6] This solution is gaining momentum, in part because it integrates with existing content management systems.

Libraries considering crowdsourcing should also look to the Australian and European library communities, as well as non-library efforts, for innovative and more seasoned examples of engaging the crowd. The National Library of Australia and the non-profit Distributed Proofreaders have organized extensive projects to correct text images scanned using OCR and enhance access by adding tags and other markup.[7] International university collaborations such as Galaxy Zoo, a Zooniverse Project, ask volunteers to classify millions of photographs of galaxies, while still other projects invite the public to upload their own artifacts and recollections for inclusion in an online collection.

The Iowa Approach

In preparation for the Civil War sesquicentennial beginning in 2011, the UI Libraries conducted a two-year reformatting project to provide comprehensive digital access to the Civil War manuscript materials in its Special Collections department, comprising approximately 50 collections containing more than 20,000 pages of correspondence and diary pages. As the scanning effort was drawing to a close, curators began to discuss ways to promote the resulting digital collection. Most of the items were handwritten and lacking transcriptions (with the exception of a small number provided by the families who donated the materials), so the idea of a transcription crowdsourcing project had strong appeal. Curatorial staff requested that the Digital Library department investigate options to develop such a project for the diaries, with the twin goals of enhancing the data by enabling full-text search of the content, and engaging the general public by allowing them to interact with the materials in new ways.

The results of initial investigations were not promising. No one in the UI Libraries possessed the programming expertise to develop a software solution from scratch. Systems staff were already so over-tasked with other library initiatives that even implementing an open-source solution was out of the question, and none of the cloud-based projects had yet emerged. Digital initiatives librarians were ready to give up and wait a few years for the technology to become more attainable when the libraries' webmaster suggested collecting submissions through the use of a web form. She wrote some simple PHP code to generate a web page that pulls diary pages from CONTENTdm, the asset management system used to host the digital collection; the page images are paired with a text box for users to type in the transcriptions; and a submit button sends an e-mail message to a departmental inbox. From there, a cataloging staff


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

John C. Brown diary, 1864

Home page 131 of 176 < Previous Page | Next Page > << First Page | Last Page >>

Sept. Saturday 24, 1864

Near East Point Ga
the now Veterans
of the 9th Iowa was
mustered out of Survis
to day but have ha
to await for our pap
to go to Headquarters
Resistered, perhaps
we may get started
for the north to moro

Sept. Sunday 25, 1864

Atlanto Ga
after a long time &
hard work & we have
got our Papers and Started
for the north. I came
in from East Point
on the Cars to day.
I met Samuel Rowe
of the 5th Iowa to day
we have got our
Transpertation ~~as~~ ready to
start as soon as there
is a train goes for the
north, we are sleeping
at the Soldiers Home

Enter transcription in the box below.
Copy the text exactly, including misspellings, abbreviations, and punctuation.
Indicate illegible text in square brackets: [?] or [town?]
Hit the return key to indicate a new paragraph.

Transcription:

Name (optional):

Email (optional):

Questions? Contact us at lib-digital@uiowa.edu

University of Iowa Libraries' Civil War Diaries & Letters Transcription Project, Transcriptionist Interface member reviews the submission, pastes it into the metadata record in CONTENTdm, and indexes the collection, at which point the transcription is live and available for searching.

This workflow has obvious drawbacks. Such a mediated system is more costly in staff time than is ideal; and the asynchronous nature of the submission process means that multiple users could work on the same page simultaneously, resulting in duplication of effort. Nevertheless, the pilot moved into the testing phase, where another issue emerged: some staff members voiced concerns about the quality of submissions, questioning whether the public was qualified to do the work. While these concerns do have some validity, the overall consensus was that some imperfect access was better than none, and any staff-side inefficiencies would be worth the trade-off in public outreach benefits.

The Civil War Diaries Transcription Project site was launched in early spring of 2011. Standard promotional tools of press releases and blog posts drew a little attention, but the crowds needed to drive the project proved to be elusive, so the next two months were spent focusing on promotion to historians, Civil War enthusiasts, and genealogists. In early June the project was featured on the American Historical Association blog;[8] from there, it was picked up by reddit.com, a social media news site.[9] The response to the reddit post was enormous; the day it went up, web statistics jumped about 7,000 percent (from typically 1,000 users to about 70,000), and the Digital Library server crashed, remaining out of commission for the next several days until the traffic became more manageable. Since then, response has calmed down quite a bit, but the project still retains a core stable of loyal transcriptionists.

One of these core transcriptionists, Dave Hesketh, is a 69-year-old retiree in the north of England who has currently transcribed 140 pages and is still going strong. Regarding the family whose papers he has spent the most time working with, he says that they "have become almost an extended part of my own:"

> …[T]hese diaries & letters reveal the lives of "ordinary" people…, rather than those of politicians, generals, and the like (whose actions, words, and deeds are generally sanitised for public dissemination). These people come alive; you come to share their hopes, their fears, their everyday concerns—the price of food for themselves, and for their animals, the cost of

> doctor's visits, their illnesses, and, sadly, their deaths. Yes—you do become absorbed in the people, and losing one of them is like losing a relative or a friend.[10]

Next Steps

In October, after volunteer transcriptionists had made short work of the original collection of 38 diaries, the project scope expanded to include 80 sets of letters containing over 5800 pages of correspondence. With this additional content, the newly renamed Civil War Diaries and Letters Transcription Project supplements the first-hand accounts of soldiers with those of their families and friends back home, allowing a much fuller view of life during the war.

This winter, the project will expand further with the creation of the Crowdsourcing Collections @ Iowa gateway site, which will provide access to additional opportunities for users to help enhance the UI Libraries' content. Next up is crowdsourcing transcription for items outside of the Civil War materials, including manuscript cookbooks from Special Collections, as well as 19th-century children's diaries held in the UI's Iowa Women's Archives, and in a partner institution, the State Historical Society of Iowa. This spring, the libraries also plan to move a Flickr pilot into production to allow commenting and tagging for historic photographs harvested from the Iowa Digital Library, migrating the user-generated data from Flickr back to the metadata records into the Digital Library. This will follow similar workflows to those developed for the transcript project.

Conclusion

This success of the Civil War project not only confirmed public interest in such endeavors, but was evidence that crowdsourcing efforts can get off the ground without computer programmers, specialized software, or major grant funds. The highly mediated nature of our workflow can lead us to feel a little self-conscious when compared to some of the more high-tech efforts—Iowa's transcription project runs on "peopleware" rather than software—but it makes sense for the institution. Overstaffed with talented, detail-oriented catalogers in technical services and understaffed in IT, Iowa adapted its crowdsourcing plans to fit, in order to "go to war with the army we've got." In the battle for public engagement and value-added collections, the ends have more than justified the means.

[1] Government & Heritage Library, State Library of North Carolina, Transcription Project, http://www.flickr.com/photos/statelibrarync/collections/72157628080641877.

[2] The Frederick Douglass Diary—A Written Rummage Project, http://frederickdouglassdiary.wikispaces.com/.

[3] Andrew S.I.D. Lang and Joshua Rio-Ross, "Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents," *Code4Lib Journal*, no. 15 (October 31, 2011), http://journal.code4lib.org/articles/6004.

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

[5] FromThePage, http://fromthepage.com/.

[6] Papers of the War Department: 1784 to 1800, http://wardepartmentpapers.org/transcribe.php.

[7] Rose Holley, "Crowdsourcing: How and Why Should Libraries Do It?," *D-Lib Magazine*, 16, no. 3/4 (March/April 2010), http://www.dlib.org/dlib/march10/holley/03holley.html.

[8] Elisabeth Grant, "Crowdsourcing the Civil War," *AHA Today*, June 7, 2011, http://blog.historians.org/resources/1346/crowdsourcing-the-civil-war.

[9] Dick_long_wigwam, "TIL How to Participate in History while Sitting on My Ass by Transcribing Civil War Diaries Online," *reddit*, June 8, 2011, http://my.reddit.com/r/todayilearned/comments/humy3/til_how_to_participate_in_history_while_sitting.

[10] Dave Hesketh, e-mail correspondence with University of Iowa Libraries, November 15, 2011.

To cite this article: Nicole Saylor and Jen Wolfe. "Experimenting with Strategies for Crowdsourcing Manuscript Transcription." *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC*, no. 277 (December 2011): 9–14. http://publications.arl.org/rli277/.