# Transcribing the Past: Crowdsourcing Transcription of Civil War Manuscripts

*By Jacquelyn Slater Reese*

ABSTRACT: Libraries and archives are using crowdsourcing in various ways. From entering data to transcribing newspapers, it is a tool to engage users while accomplishing a goal for an organization. Crowdsourcing's many benefits and challenges should be weighed when considering this technique. This article describes the planning, execution, and lessons learned from a grant-funded crowdsourced project at the University of Oklahoma to transcribe a diary and letters in commemoration of the sesquicentennial of the US Civil War.

## Introduction

Special collections and archives often seek ways to enhance community outreach and make their collections more visible and accessible. Over the past decade, numerous institutions have established projects that utilize "the crowd," or the general public, to assist with some type of endeavor using technology. Crowdsourcing involves partnering with a distributed population via the Internet to achieve a task or solve a problem ranging from transcription to improve access to documents and videos, to image identification through sites such as Flickr.

Crowdsourcing can be used in many different ways. Learning how others have successfully employed this technique, as well as the challenges encountered, can inform individuals and institutions as they consider crowdsourcing. This article discusses Transcribing the Past: Civil War Manuscripts, a project undertaken at the University of Oklahoma Libraries, which used crowdsourcing to transcribe a series of letters and a soldier's diary from the US Civil War. The grant writing, project planning, and implementation processes will be described, and the project's results and lessons learned, along with future crowdsourcing plans, will be discussed.

## Literature Review

Jeff Howe first used the term "crowdsourcing" in a 2006 *Wired* article discussing how businesses use "the crowd" in different ways.[1] Early examples of crowdsourcing include websites such as YouTube, Flickr, and Wikipedia, with the focus on content creation rather than consumption.[2] Crowdsourcing's adaptability makes a singular definition difficult. "The crowd" is defined as the participants, and "sourcing" as the procurement practice for finding, evaluating, and engaging suppliers of services and/or goods.[3] A project's needs regarding size, heterogeneity, and knowledge determine the composition of the crowd. The work done by the crowd must have a goal and be purposeful, and tasks of varying complexity are proposed.[4] Compensation comes in multiple forms,

including financial remuneration, social recognition, or the reward of benefiting the common good. The crowdsourcer is anyone who needs a problem solved or a task completed. Participation is distributed online, and participants are solicited through a flexible, open call typically issued via the Internet.[5]

Institutions of varying sizes undertake crowdsourcing projects, and each uses this technique in a slightly different manner. Ellen Forsyth highlighted the gamification or reward element of crowdsourcing and how it can be used for learning, focusing on projects such as Old Weather and Trove newspaper transcriptions.[6] Mary Flanagan and Peter Carini emphasized that using a gamification approach can lead to more participation and, in the case of Metadata Games, more tags, than a nongaming approach.[7] Several articles profiled large, well-known crowdsourcing projects such as the Library of Congress's Flickr Commons Project, the National Archives and Records Administration's Citizen Archivist Dashboard, the Smithsonian's Transcription Center, and the New York Public Library Labs projects.[8] Other projects include the University of Louisville's *Louisville Leader* newspaper transcription project, the California Digital Newspaper Collection, and the University of Alabama's Tag It—A Historical Photograph Tagging Project.[9] *Crowdsourcing Our Cultural Heritage,* which included case studies and essays regarding cultural heritage crowdsourcing, addresses theoretical and practical benefits and challenges of this technique.[10]

Crowdsourcing has numerous benefits. Jennifer A. Bartlett highlighted how these projects can build up public engagement, foster collaboration between an organization and its users, and complete projects that lack institutional resources.[11] Meredith Schwartz emphasized that endeavors that engage a community can lead to more public buy-in, increased use, and more sustained use than a passive exhibit. Users can influence the development of tools by providing input on what features they would like to see in new versions.[12] Though the preparation and implementation stages may involve more work, Dimitra Anastasiou and Rajat Gupta argued that the task to be accomplished can sometimes be done more quickly with the power of the crowd. Monetarily, crowdsourcing is more cost-effective than outsourcing. It still takes an investment of time and money during implementation, particularly for checking transcriptions, but the long-term maintenance costs are lower than for other solutions.[13] These types of projects can also increase publicity and awareness of specific topics, collections, and types of resources. Tim Causer, Justin Tonra, and Valerie Wallace found that these projects also increase access to previously hidden sources.[14]

Crowdsourcing has many benefits, but it also poses challenges. As Matthew Lease discussed, the distributed workforce spreads the load among volunteers, but it can lead to lower quality connections between the institution and the participants. Some projects allow participants to remain anonymous, limiting the opportunities to build rapport with the institution and among the participants. Filtering out "spammers," assessing participant quality in general, and consolidating data from multiple voices—both in the case of data entry and transcription—also can be problems. Eliminating the minority voice, or assuming the majority equates to quality, can easily occur in crowdsourcing

projects.[15] Quality of translations or transcriptions has always been a concern, especially when the institution has no control over who participates. Anastasiou and Gupta noted that, in some projects, especially if participants can communicate with one another, management and control of the crowd can be a challenge when one participant takes over the project. Issues such as privacy, ownership, intellectual property, and anonymity, which vary by project, also can create problems.[16] A final challenge many projects encounter is funding. Causer, Tonra, and Wallace noted that many crowdsourcing projects begin with grant funding, yet for long-term projects to continue, other funding sources must be pursued. A typical grant period may not be enough time to finish a project.[17]

Undertaking a crowdsourcing project entails weighing potential benefits and challenges, looking at similar projects for lessons learned, and considering if crowdsourcing is the best solution. Rose Holley offered several tips for successful crowdsourcing efforts in a 2010 article, where she divided her checklist into four areas: The Thing, The System, The People, and The Content. The Thing includes having a clear goal with a big challenge, showing progress, and posting results. The System should be easy, fun, reliable, quick, intuitive, and include options. The People area includes acknowledging and rewarding high achievers, remembering it takes a team to support the project, and trusting the participants. The Content should be interesting, new, and involve a history or science topic.[18]

Crowdsourcing transcription projects develop in different ways. Some, such as the *Louisville Leader* project, use existing digital files housed in CONTENTdm with a new transcription infrastructure using Scripto and Omeka built for the crowdsourcing project.[19] Others, like DIY History at the University of Iowa, use a similar procedure, but instead of software such as Omeka and Scripto, they employ a simple web form to pull CONTENTdm digital images onto pages that pair the images with text boxes for typing the transcripts.[20] Some projects begin with previously digitized content, but transform into much larger enterprises. The New York Public Library (NYPL) Labs built on content and processes begun by the Digital Library Program in the NYPL Digital Gallery and introduced public interaction with the content through projects such as the Map Warper and What's on the Menu?[21]

Archives use crowdsourcing in several ways, including for collection description projects. Zoe D'Arcy described how the National Archives of Australia used crowdsourcing to transcribe and correct consignment lists found in archival boxes to increase access to materials.[22] The University of Michigan used collaborative cataloging to create fully cataloged records for Islamic manuscript collections, and, though the participation from the larger scholarly community was smaller than expected, the local crowd participated extensively in this project.[23]

## Background

Grants are often a way to jumpstart a project using external resources. The impetus for submitting a grant application can vary, but in this instance new library administration

leadership at the University of Oklahoma Libraries encouraged units to pursue grant opportunities. After selecting the Amigos Library Services' Fellowship and Opportunity Award Program as a possible funder, several topics were discussed and potential projects outlined. As this would be a new endeavor for the library system, the planning team chose a crowdsourced transcription project of two Civil War manuscript collections. This topic was chosen to commemorate the sesquicentennial of the largest domestic conflict in US history. The Western History Collections had yet to plan a program for this anniversary, so this project would serve as such and engage the larger community with the collections. The amount of material chosen would represent a large enough sample to determine the usefulness of the transcription tool and the process, while not being too large for the organization.

This project featured six objectives:

1. Apply Scripto for crowdsourcing.
2. Develop a compelling website to recruit and retain volunteer transcriptionists.
3. Triangulate transcribed manuscript materials.
4. Promote and make freely available unique Civil War collections in observance of the Civil War sesquicentennial.
5. Incorporate the final product into the institutional repository to make it accessible to the public and promote the special collections of the institution.
6. Serve the needs of scholarly and library communities through the project content and the development of a process that can be replicated.

## Content

The two collections selected for transcription contain very different materials. The Charles Evans Collection consists of 52 letters from Lyle Garrett, a lieutenant in the Twenty-third Iowa Volunteer Infantry, and 56 letters from his wife, Mary, totaling 610 pages. After he enlisted in September 1862, Garrett's unit traveled through Missouri, Arkansas, Texas, Louisiana, Mississippi, and Alabama. He was promoted in late 1863. Garrett wrote his wife often, describing camp life, troop movements, and attitudes toward soldiering. He also gave Mary instructions regarding managing affairs at home. His letters include thoughts on issues of the day and observations made during his travels, such as other theaters of war, the destruction caused by the war, slavery, and conditions in the South. Mary's letters to her husband contained in this collection began in February 1863, and she discussed her plans to occupy her time while he was gone, the government and the war, her husband's treatment as a soldier, and events back home.

The second selection, from the Sherry Marie Cress Collection, was the diary of Charles Kroff, who served from 1861 to 1865. When Kroff enlisted on July 12, 1861, he entered Company F of the Eleventh Indiana Volunteer Infantry. His regiment fought in 15 battles and came under fire 77 days during his four years and one month of service. His diary, kept from the day he enlisted until the day his regiment mustered out in July 1865, chronicles daily military life and includes details of the battles of Fort Donelson,

Shiloh, and Corinth, and the siege of Vicksburg. Kroff made one additional diary entry on December 11, 1909, his 72nd birthday. This diary provides a different perspective than the letters between the Garretts as it preserves one person's thoughts rather than letters written as a conversation.

## Grant Awarding and Implementation

In July 2013, two months after submitting the final grant proposal to Amigos Library Services, the team received notification of partial funding for the Civil War manuscript transcription project. The team then began preparing for project implementation, including digitizing the documents as high-resolution image files. Later-than-anticipated receipt of funding slightly delayed programming and website design, as most of the funding was allocated for these areas, and required a modification of the original time line.

During these preparation and early implementation phases, several changes within the library system affected the project. A programmer was hired to work on the institutional repository, bringing in-house technical expertise for website construction and tool design to the project. The organization also hired a website coordinator, originally named as a web design consultant in the grant proposal. Both individuals had many duties and demands on their time, however, so an outside consultant was hired to create style tile designs.

Throughout these early stages, which lasted one year, tool development remained the most difficult task. The programmer had to balance conflicting requirements within the proposal while creating the transcription tool. The proposal stated the wiki-style Scripto transcription tool would be used, but this contradicted the proposal's double-blind parallel transcription process, which required two people to transcribe each document without seeing each other's work, eliminating Scripto as a tool option. Scribe, another tool mentioned in the proposal, was created for the Old Weather project—a main inspiration for this project. Investigation, however, found that this tool is better suited for data such as those found in ships' logs, the Old Weather project's focus. Though Scribe is a parallel transcription tool, it does not have a large user community, is not being maintained, and uses Ruby on Rails, a different content management system (CMS) than Drupal, which the library development team uses. Transcribr, a Drupal-based tool with a large user base and ongoing development, uses a wiki-style transcription process, eliminating it as an option. In the end, a unique Drupal-based parallel transcription tool that uses add-ons such as the DeepZoom module was designed.

Website development also remained a focal point throughout the implementation phase. The programmer and website coordinator worked on the website following the initial style tile designs. The page for transcribing image files featured two main components. The image of the handwritten document occupied the majority of the window in the middle of the screen. The DeepZoom tool allowed for zooming in to read specific words and decipher handwriting. Below the image, a collection reference denoted either

"My Dear Mary" for the Evans Collection or "A Soldier's Diary" for the Cress Collection. A text box for entering the transcript appeared below this information. This design allowed volunteers to see the words to transcribe and their typing at the same time.

Front-page design is important in creating an engaging and usable website. The "Transcribing the Past" homepage featured a menu bar on the left including links to Home, About the Project, Transcribe, Get Help, Contact, and Login pages. The homepage featured a brief overall project description with columns for both collections. Users could easily begin transcribing from the homepage by clicking the "Start Transcribing" button beneath each collection description. The Amigos Library Services logo appeared on the homepage as the sponsoring organization.

The next-level pages are also important in making a website usable. The About the Project page featured the project's purpose, background, and general participation information. The Transcribe menu option led to the two projects, and the Get Help menu option included information about how to get started transcribing, guidelines and helpful hints, and a link to the Facebook group page. The Contact menu option led to a web form for submitting questions, and the Login option led to the login screen, where participants created a username and password when registering. Users were informed of the research study project and provided an informed consent form when registering.

Originally titled "My Dear Wife" to denote Lt. Garrett's salutation style to his wife, the project title was changed to "Transcribing the Past: Civil War Manuscripts" to include Kroff's diary. This also allowed the library system to obtain the domain name transcribe.ou.edu for future transcription projects, which could fall under the "Transcribing the Past" heading.

## Institutional Requirements

Institutional reporting requirements vary by institution. At the time of this grant proposal, the University of Oklahoma Institutional Review Board (IRB) took a strict stance on research involving people. If the research findings were to be published, as required in this proposal, and the research involved people, no matter how remote the contact, restrictions applied that greatly impacted project development. For example, no volunteer identification information could be collected, participants had to remain anonymous, and staff could not directly solicit participants. Research of this nature also invokes institutional requirements, particularly required training. This slowed website production and transcription tool design as each person involved with this project who might have access to volunteers' personal information had to complete this training before beginning work.

## Website Launch and Transcription Progress

Transcription tool and website design began in earnest in January 2014, and a beta

site launched in April 2014. After several adjustments, the full site opened in August 2014. Posters advertising the project were designed in-house and, soon after the website launched, were distributed across the University of Oklahoma campus and sent to local public libraries and historical societies. Project information was posted to the OU Western History Collections Facebook page, and the project's Facebook group page became visible to the public. A press release was distributed to local media outlets several weeks after the project's launch.

Public engagement with the project occurred quickly. Transcriptions rapidly poured in, even with mainly local publicity. The campus newspaper, the *Oklahoma Daily,* published an article about the project within a few weeks.[24] Interviews with the *Norman Transcript,* the local newspaper, took place around the same time, but the article did not appear until several months later. Unfortunately, the *Transcript* article appeared just as the final transcriptions were completed.[25] Throughout the project, staff posted project updates on the Facebook group page.

Technical updates and questions were resolved quickly throughout the project. Security updates were handled manually at the project's outset, but, by its end, automatic updating was established. The developer addressed any user-submitted issues as quickly as possible. Technical questions relating to the transcription process were also answered in a timely manner. Though project staff could not directly contact volunteers, several people self-identified by contacting staff with questions regarding formatting and wanting to discuss their experiences. Several volunteers requested feedback on transcription quality and accuracy.

The transcription process progressed far more quickly than expected. Rather than taking more than a year to transcribe all the documents twice, as the proposal stated, it only took three months. Much of this was due to "super transcribers," a few volunteers who transcribed much more content than the others. The transcription project ended in November 2014. Several users who contacted project staff asking to participate after the project closed were referred to similar projects at other institutions.

Once the crowdsourcing portion of the project was completed, transcription triangulation began. As stated in the grant proposal, each page was transcribed twice in a double-blind, parallel transcription style. Triangulation and reconciliation of the two transcripts had to occur before a final transcript could be posted online. Project staff received the text files for reconciliation in January 2015, and the triangulation process lasted until March. While still performing regular job duties, one staff member reconciled the transcripts using Juxta Commons, a tool that allows for the comparison and collation of textual work versions. This tool highlights every inconsistency between two text files. The staff member used it for the first third of the transcripts, but thereafter relied more heavily on visual comparison with the original handwritten document to create a final transcript.

## Results and Outcomes

Although the launch date of the website was one year later than proposed in the grant application, all transcriptions were completed by the grant deadline, thanks to the super transcribers. One person completed 763 pages, almost half the total number of transcripts. One volunteer transcribed 164 pages, another volunteer did 75 pages, and two volunteers completed between 55 and 65 pages. Eight volunteers transcribed between 20 and 40 pages; 6 people did 10 to 20 pages; and 19 people did 3 to 9 pages. Eleven volunteers did 2 pages, and 24 people did only 1 page of transcription. Seventy-nine of the 152 registered users never transcribed any pages (see figure 1). The final number of transcripts submitted was 1,596, though this included several blank pages.

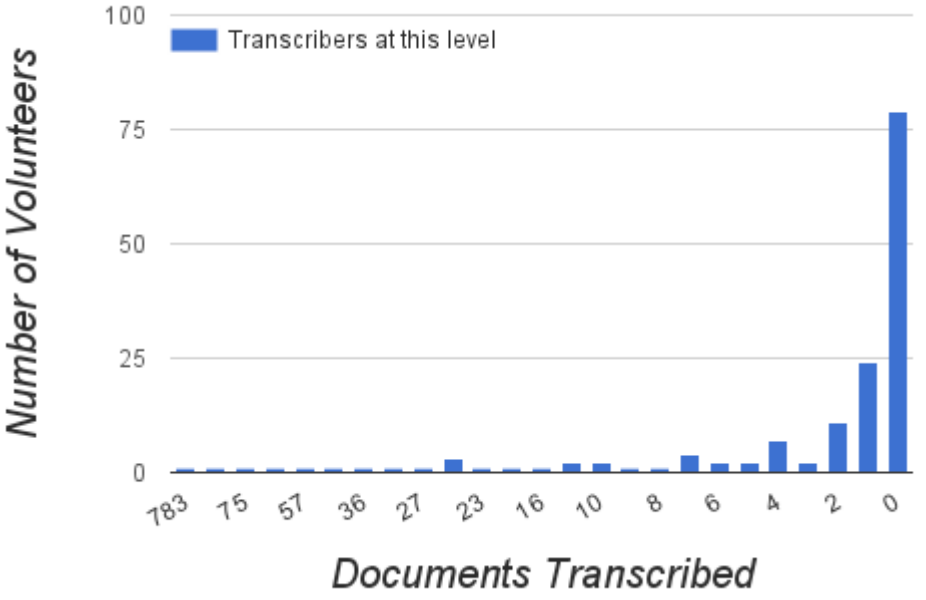


*Figure 1. This graph shows how many volunteers transcribed how many documents.*

Vandalism or junk transcripts can be a concern in crowdsourcing projects, but this project received none. Although the transcription tool did not allow users to save a transcript for later work, only 10 partial transcripts were received. These partial transcripts counted toward the total number.

This project developed much differently than anticipated, but still ended as desired. Two historically significant manuscript collections containing 787 handwritten pages now have transcripts, and 152 volunteers engaged in crowdsourcing. With this number of volunteers, consistency and quality of transcripts varied greatly. Several factors could have led to this variation, such as document content about unfamiliar Civil War events and places, the language of the day, and military terminology. Difficulty reading the documents also could

have influenced quality, whether due to faded ink, cramped handwriting, or unfamiliar cursive handwriting (see figure 2). Overall, though, most volunteers seemed conscientious and deliberate with their transcribing (see figure 3).

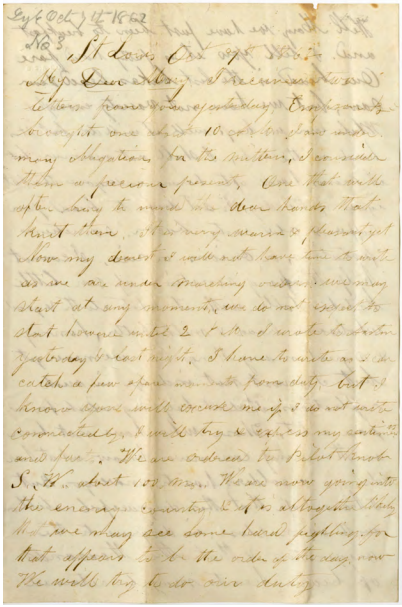

*Figure 2. This letter from Lt. Lyle Garrett is an example of faded ink and handwriting that is difficult to read. Charles Evans Collection, Box 1 Folder 2, Western History Collections, University of Oklahoma Libraries, Norman, Oklahoma.*

St Louis Oct 7th 1862

My Dear Mary, I received two letters from you yesterday, Empson B. brought one about 10 A.M. I am under many obligations for the mittens, I consider them a precious present, one that will often bring to mind the dear hands that knit them. It is very warm & pleasant yet Now my dearest I will not have time to write as we are under marching orders. we may start at any moment, we do not expect to start however until 2 P.M. I wrote to Austin yesterday & last night. I have to write as I can catch a few spare moments from duty, but I know you will excuse me if I do not write connectedly. I will try & express my sentiments and facts. We are ordered to Pilot Knob S.W. about 100 ms. We are now going into the enemy country, & it is altogether likely that we may see some hard fighting for that appears to be the order of the day now.

We will try to do our duty

*Figure 3. Transcript for letter in figure 2*

Improved access is an important result of this project. These collections were previously only available as handwritten documents. The Cress Collection containing the Kroff diary had been digitized, but the Evans Collection containing the Garrett correspondence had not. Now, both collections are digitized as preservation-quality images, and each page is transcribed. The transcripts will make these collections accessible to a wider audience, especially since the image and text files are freely available online.

The grant proposal stated six objectives. Though the objectives were met in a slightly different manner than originally envisioned, each was fulfilled. The staff modified the first objective—apply Scripto to crowdsourcing—and instead created a custom tool that worked with the library system's existing Drupal CMS and included a Drupal module to manage transcript creation. Other Drupal add-ons were used to create the site and the transcription tool.

The tool design tied closely into the second objective: develop a compelling project website to recruit and retain volunteer transcriptionists. Some original elements planned included an attractive, eye-catching homepage design and a forum for participant interaction. However, the proposal stated that partial funding, as was awarded, would result in some truncation of the website. This resulted in the website containing only the essential features—the images of the documents to be transcribed, a FAQ page, the transcription tool, user surveys, general information about the project, and required IRB documentation. Though the website did not contain the proposed user forum, it still met all the project's needs.

User engagement with a website can occur at different levels. In this instance, several very active volunteers were highly engaged with the content. One way to extend this interaction to other participants, and a common feature on similar project websites, is a user forum. However, IRB restrictions eliminated this feature. As an alternative, the staff created a Facebook page and a link to it on the website. Though many transcription projects have active user groups, this project did not generate much engagement among users. Project staff posted to the Facebook group and users commented on these posts, but volunteers did not use the page to engage each other. More user-to-user interaction might have resulted if the forum had been hosted on the project website, as some people may not have had Facebook accounts or did not want to navigate out of the project website. User satisfaction surveys can also engage users. Of the 152 registered users, only 9 people completed the survey. These users expressed overall satisfaction with the website, but they did ask for more navigation options. They also desired the ability to edit their work after submitting it. This feedback will be helpful in planning future transcription projects.

The third objective, triangulate transcribed manuscript materials, was met in a very time-intensive way. Editing transcripts is a time-consuming process, but triangulating two transcripts to create the best overall transcript of a document is even more so. Juxta Commons was used for the first one-third of transcripts, and the text files were then visually compared with the image files to determine which transcript was more accurate. Corrections were made, if needed, to create a final transcript. Eventually, the staff member performing the triangulation relied less on Juxta Commons and more on visual comparison to create the final transcript.

Though project staff knew this could be an intensive process, it was more so than anticipated because of the various ways volunteers transcribed. Not all volunteers adhered to the guidelines provided. Abbreviations were spelled out in some instances and left abbreviated in others; originally misspelled words were corrected and notated, corrected and not notated, or left misspelled; and, sometimes, words were simply improperly transcribed (see figures 4 and 5). All of these factors meant that transcript triangulation took almost as much time as the transcription process, making this the most intensive portion of the project and an unsustainable model for future projects.

The fourth objective, promote and make freely available unique Civil War collections

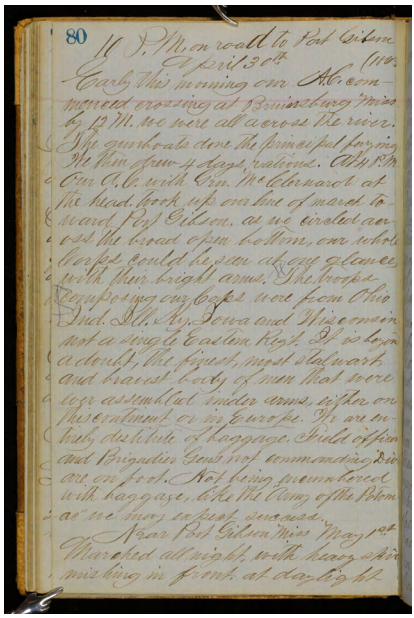

*Figure 4. This page from Charles Kroff's diary demonstrates military terminology and abbreviations. Sherry Marie Cress Collection, Folder 6, page 80, Western History Collections, University of Oklahoma Libraries, Norman, Oklahoma.*

*10 P.M. on road to Port Gibson*

*April 30th (1863*

*Early this morning our A.C. commenced crossing at Bruinsburg Miss. by 12 M. we were all across the river. The gunboats done the principal ferrying. We then drew 4 days rations. At 4 P.M. Our A.C. with Gen. McClernard at the head, took up our line of march toward Port Gibson. as we circled across the broad open bottom, our whole Corps could be seen at one glance, with their bright arms. The troops composing our Corps were from Ohio Ind. Ill. Ky. Iowa and Wisconsin not a single Eastern Reg't. It is beyond a doubt, the finest, most stalwart, and bravest body of men that were ever assembled under arms, either on this continent or in Europe. We are entirely destitute of baggage. Field officers and Brigadier Gen's not commanding Div's are on foot. Not being encumbered with baggage, like the "Army of the Potomac" we may expect success.*

*Near Port Gibson Miss May 1st*

*Marched all night, with heavy skirmishing in front. at daylight*

*Figure 5. Transcript for diary page in figure 4*

in observance of the Civil War's sesquicentennial, was easily and clearly met. From the moment the transcription site opened to the public, the collections were freely available. Both the image and transcript files are available through the University of Oklahoma's institutional repository, SHAREOK.[27] Promotion of the collections began shortly after the website's launch with publicity distributed through e-mail listservs (no direct e-mail solicitations were sent in accordance with IRB requirements), local physical postings, and media coverage. A press release describing the project led to articles in the student and local newspapers. The *Southwestern Archivist* also published an article.[28] These articles generated a surprising level of interest from the public about this project, other related collections, and OU's special collections in general. Interest in the project continued well after its completion, proving a clear demand from the public for projects that allow people to engage with historic documents and be part of something with lasting value.

The fifth objective was to incorporate the final product in the institutional repository to make it accessible to the public and promote the special collections. Although no repository existed at the time of the grant application, this objective was met in part because the organization hired the developer who worked extensively on the transcription

project to develop the repository. Having the repository in mind when creating the project's website and transcription tool made it easier to upload and incorporate these files into the repository. The scanned diary and letter pages and the accompanying transcripts were uploaded into SHAREOK in May 2015. There are no access restrictions on this website. The SHAREOK page for the project contains zip files of the high-resolution images compiled into one file per collection. Individual transcript text files and compiled transcript PDF files for each collection are also available. The PDF compilations include background information about the collections, and a readme file provides information about the project and collection summaries. At this time, the image files and transcripts are not displayed simultaneously.

The completion of transcripts for all documents, digitization of original documents, and availability of the digitized documents and transcripts through the institutional repository met the first part of the final objective—serve the needs of the scholarly and library communities through the project content. Having these documents freely available opens these collections to more in-depth research and allows comparison with other readily available sources. This also benefits the library community by adding to the existing body of freely available primary sources, which are increasingly used at all educational levels.

Several factors complicated the second portion of this objective—serve the needs of the scholarly and library communities through the development of a process that can be replicated. First, the proposal specified the use of parallel transcriptions, eliminating a wiki-style tool such as Scripto. Scribe, the best parallel transcription tool available, did not meet all the project's needs. Institutional requirements regarding participant anonymity also limited the transcription tool and website design. A custom transcription tool specific to the project was created to meet the proposal and institutional requirements, and it will not likely be usable in another situation. Institutions that use Drupal may find this project's custom tool useful, but others will need a tool specific to their CMS. Still, the general process for creating a custom transcription tool could be followed if such a tool is needed. The transcription tool code is available online.[29]

## Lessons Learned

The project implementation team learned valuable lessons throughout this endeavor. First, when composing a funding proposal, be less specific regarding the technology to be used. From the time a proposal is submitted until funding is approved, technology can drastically change. This proposal mentioned two transcription tools, neither of which met IRB requirements. Specifying parallel transcription in the proposal excluded wiki-style tools, although the proposal mentioned a wiki-style tool. However, the need to protect volunteer anonymity eliminated both tools specified in the proposal. Using a general statement regarding transcription tools would have benefited this proposal.

A second lesson learned is that including technical and subject expertise at the beginning of the proposal writing process and throughout the project is critical. In this

instance, key technical experts were not in place during the early proposal writing stage, which led to contradictions within the proposal regarding the transcription process and the tool to be used. This also led to an unrealistic time line of three months for website construction, digitization, and tool development. The time line also did not account for a delayed funding date. Once technical experts were brought in, website design and tool development became much easier.

More subject expertise during the proposal writing process could have informed material handling processes and other decisions. The assistant curator, who oversaw the Manuscript Division of the Western History Collections, was on sabbatical and unavailable during initial project planning. Though other special collections staff were involved in the early stages, they did not have the curatorial knowledge of the manuscript collections involved, which complicated work flows once the project began. Having this additional expertise involved from the beginning would have streamlined the collection side of the project.

Limitations from outside the project also affected its design and success. The IRB requirements presented unexpected external limiting factors. The team learned the IRB would review the proposal and the Office of Research Services (ORS) would submit the final proposal shortly before the final submission deadline. Knowing the potential IRB limitations from the outset would have informed the tool and project design specified in the proposal. Privacy restrictions meant project staff could not respond to forgotten password inquiries, and no direct contact in person or via e-mail could be made to recruit volunteers. Volunteer-requested qualitative feedback regarding transcriptions could not be provided as individual transcriptions had no identifying information associated with them. This limitation greatly influenced design of both the website and the transcription tool, eliminating certain website features and resulting in a custom transcription tool so specific that it will be difficult for other institutions, or even OU Libraries, to use again. The unanticipated IRB-required training, though necessary for projects involving contact with volunteers' personal information, hindered the technical end of this project. Not every institution will face the same IRB requirements, but being aware that this could occur when submitting a proposal must be considered.

The team also learned that a better transcription reconciliation process will be needed for future projects. Requiring parallel transcription and triangulation of transcripts into a final version was time intensive, requiring the full-time attention of one staff member for three months. To scale up this type of project, a different method would be needed. Some projects rely on graduate assistants or the participants to perform the editing. A graduate assistant requires funding, however, and that funding is often written into the grant proposal. This project did not use funds for a graduate assistant. With wiki-style transcription tools, volunteers can edit each other, sometimes with different levels of volunteers performing different tasks. If volunteers made the majority of editing changes, it would take less time for staff to evaluate a transcript for final online publication.

This project's popularity is a final positive lesson. Staff anticipated interest from certain

user groups, but they were thoroughly surprised by the project's popularity. Instead of one year, as the original time line stated, it only took 81 days to complete almost 1,600 transcripts. Even with somewhat limited publicity, word spread quickly, and regular volunteers became dedicated transcribers. Though not all volunteers were super transcribers, enough people devoted themselves to the project to enable its rapid completion. The number of people who signed up near the project's completion also testifies to its popularity.

An important element of this popularity was the continuing demand to participate. Institutions should be ready with new projects once one is completed. In this instance, other projects were not ready, so interested volunteers were referred to other institutions. Much goodwill was created with this project, and capitalizing on that with another project would have benefited the institution.

## Future Projects

Rather than a project-specific website, the general domain transcribe.ou.edu was created so future transcription projects could live at the same site, where multiple projects might run simultaneously. Multiple special collections within the institution have numerous potential transcription projects. The institution's digital collections already contain thousands of pages of handwritten documents regarding Native Americans and early manuscripts about the history of science. Many of these documents and manuscripts could easily be imported into the transcription site. Any future projects would rely on a wiki-style transcription tool such as Transcribr, a Drupal-based tool that would easily fit into the existing CMS. Add-ons would allow customization, but would not require building a tool from scratch. Using the wiki-style transcription tool would aid the transcription reconciliation process, along with engaging volunteers at different levels to perform various editing tasks. All files would be deposited in the institutional repository, as were those in this project.

## Conclusion

Crowdsourcing has many different potential uses in libraries and archives. Whether it is entering data or transcribing handwritten documents, leveraging a distributed workforce can save organizations time and money. Projects can highlight important historical events and encourage interaction between an organization and its community, as is evidenced by the popularity of the Transcribing the Past: Civil War Manuscripts project at OU Libraries. Whenever an organization undertakes such a project, other institutions can learn from the experience. Involving technical and subject experts during the grant-writing process, being flexible when describing technology to be used, gaining awareness of institutional limitations, and having future projects ready when one project is completed are all beneficial lessons for institutions considering such a project. Technical benefits of working with this type of technology include sharing coding and experiences using different transcription tools with other developers. As libraries and archives look for ways to engage their user communities, they should continue to explore crowdsourcing as an avenue to achieve this goal in new and unique ways.

## ABOUT THE AUTHOR

Jacquelyn Slater Reese is the Western History Collections librarian and assistant professor of bibliography at the University of Oklahoma Libraries. She earned her BA in history at Pittsburg State University and her MLIS at the University of Oklahoma.

## NOTES

1. Jeff Howe, "The Rise of Crowdsourcing," *Wired* 14, no. 6 (2006), accessed September 29, 2015, www.wired.com/2006/06/crowds.
2. Bernardo A. Huberman, Daniel M. Romero, and Fang Wu, "Crowdsourcing, Attention, and Productivity," *Journal of Information Science* 35, no. 6 (2009): 758, accessed February 12, 2013, DOI: 10.1177/0165551509346786.
3. Enrique Estellés-Arolas and Fernando Gonzalez-Ladrón-de-Guevara, "Towards an Integrated Crowdsourcing Definition," *Journal of Information Science* 38, no. 2 (2012): 189, accessed February 12, 2013, DOI 10.1177/0165551512437638.
4. Ibid., 193–94.
5. Ibid., 195–96.
6. Ellen Forsyth. "Learning through Play: Games and Crowdsourcing for Adult Education," *Aplis* 25, no. 4 (2012): 169–70.
7. Mary Flanagan and Peter Carini, "How Games Can Help Us Access and Understand Archival Images," *The American Archivist* 75 (Fall/Winter 2012): 532.
8. Jan Zastrow, "The Digital Archivist. Crowdsourcing Cultural Heritage: 'Citizen Archivists' for the Future," *Computers in Libraries* 34, no. 8 (2014): 22–23.
9. Jennifer A. Bartlett, "Internet Reviews: Crowdsourcing in Libraries and Archives," *Kentucky Libraries* 78, no. 2 (2014): 7.
10. Mia Ridge, ed., *Crowdsourcing Our Cultural Heritage* (Burlington, VT: Ashgate Publishing Group, 2014).
11. Bartlett, "Internet Reviews," 8.
12. Meredith Schwartz, "Dicing Data at NYPL Labs," *Library Journal* 137, no. 14 (2012): 23.
13. Dimitra Anastasiou and Rajat Gupta, "Comparison of Crowdsourcing Translation with Machine Translation," *Journal of Information Science* 37, no. 6 (2011): 639–41, accessed February 12, 2013, DOI 10.1177/0165551511418760.
14. Tim Causer, Justin Tonra, and Valerie Wallace, "Transcription Maximized; Expense Minimized? Crowdsourcing and Editing The Collected Works of Jeremy Bentham," *Literary and Linguistic Computing* 27, no. 2 (2012): 132–33, accessed June 11, 2015, DOI 10.1093/llc/fqs004.
15. Matthew Lease, "On Quality Control and Machine Learning in Crowdsourcing," *Proceedings of the 3rd Human Computation (HCOMP) Workshop at the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI)* (2011): 97–99.
16. Anastasiou and Gupta, "Comparison of Crowdsourcing Translation with Machine Translation," 641–42.
17. Causer et al., "Transcription Maximized; Expense Minimized?," 132.
18. Rose Holley, "Crowdsourcing: How and Why Should Libraries Do It?, *D-Lib Magazine* 16, nos. 3–4 (2010), accessed February 12, 2013, www.dlib.org/dlib/march10/holley/03holley.html or DOI:10.1045/march2010-holley.
19. Caroline Daniels, Terri L. Holtze, Rachel I. Howard, and Randy Kuehn, "Community as Resource: Crowdsourcing Transcription of an Historic Newspaper," *Journal of Electronic Resources Librarianship* 26, no. 1 (March 2014): 39–40, accessed June 11, 2015, DOI 10.1080/1941126X.2014.877332.

20. Nicole Saylor and Jen Wolfe, "Experimenting with Strategies for Crowdsourcing Manuscript Transcription," *Research Library Issues: A Quarterly Report from ARL, CNI, and SPARC* 277 (December 2011): 10–12.
21. Ben Vershbow, "NYPL Labs: Hacking the Library," *Journal of Library Administration* 53, no. 1 (2013): 80–89, accessed June 11, 2015, DOI 10.1080/01930826.2013.756701.
22. Zoe D'Arcy, "'The Hive': Crowdsourcing the Description of Collections," in *Description: Innovative Practices for Archives and Special Collections,* ed. Kate Theimer (Lanham, MD: Rowman and Littlefield, 2014).
23. Evyn Kropf, "Collaboration in Cataloging: Sourcing Knowledge from Near and Far for a Challenging Collection," in *Description: Innovative Practices for Archives and Special Collections.*
24. Daisy Creager, "Calling All Scholars, History Buffs, and Transcribers," *Oklahoma Daily,* September 3, 2014, accessed April 13, 2016, www.oudaily.com/news/calling-all-scholars-history-buffs-and-transcribers/article_a649b9f0-32e8-11e4-87d7-0017a43b2370.html.
25. Jerri Culpepper, "Transcribing the War," *The Norman Transcript,* October 31, 2014, accessed April 13, 2016, www.normantranscript.com/news/transcribing-the-war/article_90eaef20-6119-11e4-9062-73b03a985955.html.
26. "Transcribing the Past Project Data," SHAREOK, accessed April 13, 2016, shareok.org/handle/11244/14633.
27. Kristina Southwell, "OU Libraries Launches Transcription Project for Civil War-Era Documents," *Southwestern Archivist* (November 2014), accessed April 13, 2016, www.southwestarchivists.org/resources/Documents/Newsletters/SWA2014_v37_Nov2014.pdf.
28. "OU Libraries/mdw," GitHub, accessed April 30, 2016, https://github.com/OULibraries/mdw.