3-11-2014

Community as resource: crowdsourcing transcription of an historic newspaper.

Caroline Daniels
University of Louisville, c0dani01@louisville.edu

Terri L. Holtze
University of Louisville, terri.holtze@louisville.edu

Rachel I. Howard
University of Louisville, rachel.howard@louisville.edu

Randy Kuehn
University of Louisville, rtkueh01@louisville.edu

Original Publication Information

Howard, Rachel I., Caroline Daniels, Terri Holtze, and Randy Kuehn. "Community as Resource: Crowdsourcing Transcription of an Historic Newspaper." *Journal of Electronic Resource Librarianship* 26(1) (2014): 36-48.

Community as Resource: Crowdsourcing Transcription of an Historic Newspaper

Caroline Daniels
Terri L. Holtze
Rachel I. Howard
Randy Kuehn

Abstract. Like many cultural heritage institutions, the Archives and Special Collections at the University of Louisville faces the dichotomy of material abundance and budgetary scarcity. Driven by the desire to make historical primary sources accessible online, this organization harnessed the power of the public to transcribe the *Louisville Leader*, an historic African American newspaper. The first sections of this article define crowdsourcing and describe how it was implemented at the University of Louisville, including the tools adopted and the process used. The latter sections outline the marketing strategy, the public response, and lessons learned from this ongoing project.

Keywords. Crowdsourcing; Historical newspapers; Digitization; Digital collections

Refer correspondence to Terri L. Holtze, University of Louisville Libraries, Louisville, KY 40292. E-mail: terri.holtze@louisville.edu

Received: July 1, 2013
Accepted: August 1, 2013

Community as Resource: Crowdsourcing Transcription of an Historic Newspaper

Libraries and archives around the globe contain vast stores of resources originally created in print formats. The information and insight contained in those documents is only as accessible as the documents themselves, so a major movement is underway to digitize and electronically disseminate these items. Unfortunately, the paucity of resources in terms of funding and staffing at these cultural heritage institutions does not match the richness of the materials they hold, so they explore options to do what they can with what they have. Within these constraints the Archives and Special Collections (ASC) of the University of Louisville Libraries (ULL) embarked on an experiment to harness some of the free time of the public in the interest of preserving and promulgating the contents of an historical African American newspaper called the *Louisville Leader*.

This article defines crowdsourcing and describes the process used at the University of Louisville to set up a system and mechanism for the public transcription of newspaper articles. Details include the criteria for choosing the software used – Omeka, Scripto, and Zoom.it – as well as how the tools were set up and the articles processed. This article details some of the methods used to bring this project to the attention of the public, including an extended marketing plan.

Finally, the project's planning team shares their insights on the impact of this crowdsourcing project. They discuss problems encountered, user interaction, and the benefits of crowdsourcing above and beyond the articles transcribed. In the end the authors answer the question, "Was the outcome of the crowdsourcing project worth the effort?"

Literature Review

Jeff Howe of *Wired* magazine is considered to have originated the term "crowdsourcing" in 2006. The term refers to the use of members of the public, rather than employees or contractors, to accomplish

tasks (Howe, 2006). Subsequent refinements of the definition have emphasized that the crowd is "employed" using Internet technologies (Saxton, Oh & Kishore, 2013, pp. 2, 4). Crowdsourcing takes a variety of forms, and the individuals performing the work may be unpaid, or they may be paid on a piece-rate basis. In some cases, individuals enter a contest, submitting their solution to a problem, be it coding in software or other kinds of design**.** While some definitions include projects such as Wikipedia (Howe, 2008, pp. 57-61), others argue that true crowdsourcing is directed by the organization rather than the crowd (Brabham, 2013, p. xxi).

The bulk of the literature on crowdsourcing focuses on business applications, but an increasing number of archives, museums, libraries, and related institutions and organizations are testing the usefulness of crowdsourcing to accomplish tasks that they lack either the time or the expertise to complete. For example, the National Institute of Standards and Technology turned to the crowd for assistance in identifying scientific instruments (http://nistdigitalarchives.contentdm.oclc.org/), and the Citizen Science Alliance is calling on the public to classify galaxies (http://zoo2.galaxyzoo.org/). Other efforts are devoted to identifying photographs, such as Dartmouth College's project (Howard, 2011); still others, such as University College London's *Transcribe Bentham* project (http://www.ucl.ac.uk/Bentham-Project/transcribe_bentham), focus on transcribing and marking up manuscript materials. While naming obscure or esoteric technology falls in the realm of specialists, many topics and tasks – such as transcribing – can be approached by a wide variety of individuals.

Other projects have highlighted the concerns that come with involving the public in such work. First, there may be doubts as to the quality of work that unknown, untested workers will produce (Causer, Tonra, & Wallace, 2012, p. 121; Lang & Rio-Ross, 2011; Saylor & Wolfe, 2011). After all, if the work is of very poor quality, it may not matter that it is free. Thus, some projects incorporate a proofreading phase. *Written Rummage*, a transcription project involving one of Frederick Douglass's journals, used

Amazon's Mechanical Turk, a fee-based, piece-rate service, to transcribe and then proofread the content. In order to foil workers who might attempt to game the system, the project's directors inserted intentional errors into the transcribed text so they could test whether the proofreaders had done their work (Lang & Rio-Ross, 2011). In other cases, such as *Transcribe Bentham* and New York Public Library's (NYPL) *What's on the Menu?*, institutional staff members perform quality control (http://menus.nypl.org/). While the public does initial transcribing in *What's on the Menu?*, and users can review what others have transcribed, NYPL staff do another round of review. In addition to providing some form of mediation, some projects have concluded that imperfect transcriptions are better than none at all (Saylor & Wolfe, 2011, p. 12).

Another concern that surfaces with unpaid crowdsourcing projects is how to motivate participants. The University College London, for example, feared that participation would decline if they did not acknowledge volunteers' contributions directly – and quickly (Causer, Tonra, & Wallace, 2012). The solution often takes the form of recognizing "super-contributors." For example, the California Digital Newspaper Collection lists users with the "top text correctors" on its front page (http://cdnc.ucr.edu/cdnc), while Zooniverse's Old Weather Project (http://www.oldweather.org/) allows participants to move up the ranks from cadet to captain. Others embed their tasks in games in order to add interest to what can be tedious work. For example, Dartmouth College used Metadata Games, in which pairs of "players" try to come up with the same tags for images, to encourage people to identify pictures in Dartmouth's collections (Howard, 2011). A third area of perceived risk is a result of the potential anonymity of the crowd: spamming and the submission of spurious, offensive, or intentionally incorrect content (Causer, Tonra, & Wallace, 2012). Many projects thus require a login (this also assists with tracking users' contributions for rewards); proofreading processes are also designed to help weed out undesirable content.

Transcribing archival materials is an example of work that is within the reach of most adult users of the Internet. Several projects invite the public to transcribe handwritten letters, diaries, and the like, as well as items such as menus that have script-like fonts. Examples include the Iowa DIY History Project (http://diyhistory.lib.uiowa.edu/), NYPL's *What's on the Menu?,* and the National Archives and Records Administration (NARA) Pilot Project (http://transcribe.archives.gov/). Because optical character recognition (OCR) software cannot recognize handwriting, the only way to make these items full-text searchable is to transcribe them. This is time-consuming if done in-house, and many cultural memory institutions do not have funds to outsource this work. The crowd, therefore, becomes an enticing potential source of assistance. While the NYPL and NARA developed one-off software to support their crowdsourcing projects, the University of Iowa initially used a simple web form and CONTENTdm (Saylor & Wolfe, 2011). The Iowa project establishes a process that many institutions can follow, even if they lack programming staff.

OCR proofreading or complete transcription can greatly enhance the searchability of newspaper texts, and crowdsourcing has been utilized to this end (Zarndt & Geiger, 2013). While newspapers, printed rather than handwritten, are relatively good candidates for OCR, they do present problems. The complex layout, involving columns and jump pages (when an article is continued from one page to another non-sequential page), can challenge OCR software; older newspapers may have archaic letter forms as well (Tanner, Munoz, & Ros, 2009). In response to less-than-ideal OCR results, several institutions have initiated text-correction projects, including the National Library of Australia's Trove (http://trove.nla.gov.au/; Holley 2009) and the California Digital Newspaper Collection, as well as other sites that use Veridian software. Both the Trove system and Veridian support the display and correction of individual articles. However, the Trove system was developed specifically for the National Library of Australia, and Veridian is a proprietary, standalone system. Despite the fact that many archives and libraries use CONTENTdm to provide access to newspapers, the literature lacks any descriptions of

projects incorporating crowdsourcing and newspaper transcription or correction for use in a CONTENTdm environment. This article seeks to fill that gap.

The *Louisville Leader* Transcription Project

The *Louisville Leader* was started by I. Willis Cole in 1917 as a Black-community newspaper in the fullest sense: it covered local religious, educational, social, fraternal, and sporting activities, as well as national and international news. The *Leader* announced births and deaths, named those suffering from illness, listed Louisville churches and their schedules of services, and printed news items from Black correspondents elsewhere in the state. The eight-page weekly advertised black businesses and professionals and sponsored contests. The *Leader* served as a voice for civil rights, imploring Blacks to vote, and opposing Jim Crow laws, segregation, and lynching. After Cole died on February 19, 1950, his family tried to continue the newspaper, but suspended publication in the fall of 1950.

The original copies of the newspaper did not fare well once publication ceased. Initially housed in the Cole Publishing Company building, they were badly damaged by a fire in 1954. Eventually, the family gave the badly deteriorated remains of the collection to Kentucky State University in Frankfort, Kentucky, who loaned them to the University of Louisville in 1978 for microfilming by the University Archives and Records Center (UARC), now a division of ASC (http://louisville.edu/library/archives). The editor's widow supplemented this collection with additional personal copies. In the interim between 1978 and 2011, the originals completely deteriorated and are now lost. In the transition from newsprint to microfilm, colorful elements such as red headlines were lost, and the aspect ratio of the pages was modified, but it is thanks to the microfilm that a significant run of the paper exists at all.

In an effort to enhance access to this resource, and with the family's permission, the archives outsourced the scanning of the microfilm as 400 ppi grayscale TIFFs in late 2011, with an eye to including the paper in their CONTENTdm-based Digital Collections (http://digital.library.louisville.edu).

The deliverables, arranged by microfilm roll, were converted to PDFs with optical character recognition text using Adobe Acrobat. The OCR proved woefully inadequate. While software exists, such as Veridian, that can handle article segmentation, zoning (instructing the OCR reader to read down columns rather than across them), and threading (connecting the first part of an article with its continuation), no funds were available to pursue this. In the case of the *Leader*, the original newspapers were particularly worn (and often torn, burned, or damaged by smoke or water); this was compounded by the fact that the scans were not performed on the originals, but on microfilmed versions. While not surprising, the poor quality of the OCR was disappointing given the historical and genealogical value of the content.

Beginning in fall 2012, a graduate student intern began preparing the *Leader* collection for online access by separating the PDFs into files by issue and creating issue-level metadata (title, volume and issue number, date, and brief description) for upload into CONTENTdm. The software automated the splitting of the PDF into the separate pages of a compound object, which meant that a text search would bring researchers to the appropriate page within an issue rather to the first page of the issue. However, since the OCR did not make the text as searchable as hoped, the archivists began considering crowdsourcing as a solution.

Building the Transcription Infrastructure

The decision to crowdsource the transcription of the *Louisville Leader* precipitated the need to identify, evaluate, and choose software that would meet the project's requirements. The next phase of the project required the most technological knowledge: setting up the server and customizing the software displays. The planning group consisted of the Director of Archives and Special Collections, who initiated the digitization and crowdsourcing of the newspaper; the Digital Initiatives Librarian, responsible for providing online access to the materials; the Head of Web Services, whose design skills were already invoked for CONTENTdm and would also be required for transcription software; and the

Digital Technologies Systems Librarian, who brought programming and system administration skills. This group made decisions collaboratively throughout the planning process.

Fortunately, projects by other institutions provided examples to follow. To get started, the planning group reviewed several existing archival crowdsourcing projects, including the University of Iowa's *Civil War Diaries & Letters Transcription Project*, *Transcribe Bentham*, and the National Archives Transcription Pilot Project. The group investigated ten tools and projects, using a spreadsheet to track findings on the features, institutional users, and CONTENTdm compatibility of each tool. They also noted whether the software was available under an open-source license, how much development had been done on the software, what customization options were available, and how easy the implementation would be. Members of the group tested each of them from the perspective of the end-user.

In examining each project and tool, the planning group was able to evaluate different approaches to presenting the content to be transcribed, as well as the applications used. How were the newspapers presented to the public – as pages, articles, or even individual lines? How much mediation occurred to make the transcription publicly discoverable? What level of metadata was visible at the transcription stage? How were sections ready for transcription displayed to the public? What image formats were used? Did format type affect load time?

The ideal tool would have automated the process of moving the transcription into CONTENTdm. None of the products available at the time had that capability; however, one product – Scripto – did indicate that it had a CONTENTdm connector script "coming soon." This note later disappeared from the Scripto site.

Nevertheless, Scripto rose to the top of the tools list. The Roy Rosenzweig Center for History and New Media developed Scripto and made it available under an open source license (http://scripto.org/about). Because there was little funding for this project, the fact that Scripto was available without a fee was a

significant attraction. In addition, Scripto was designed to be relatively easy to implement. The software is compatible with WordPress, Drupal, and Omeka, which was also developed by the Rosenzweig Center. The ULL had adopted Omeka in 2010 to display its Digital Exhibits (http://exhibits.library.louisville.edu/omeka/). The familiarity factor made Scripto particularly appealing: Since Omeka had already been running successfully on a production server for nearly two years, the Office of Libraries Technology, the Head of Web Services, and the Digital Technologies Systems Librarian had working knowledge of the back end of the product. The base installation offers a robust set of options and numerous plugins offer many additional features, which had served the ULL's past Omeka projects quite well. For the purpose of the transcription project, however, the planning group would have to delve deeper into the software code in order to make changes to both the functionality and interface to attain the goal of providing an end-user experience that would be straightforward and fulfilling.

Once the decision was made to use Scripto as the project's transcription tool, the Digital Technologies Systems Librarian worked with the Office of Libraries Technology to plan for the initial setup of a virtual test server with the prerequisite software components for the project. The Linux System Administrator created a base system that included the installation of the following software packages: Linux 2.6.32 (CentOS), Apache 2.2.15, PHP 5.3.3, Omeka 1.5.3, MediaWiki 1.19.2, Scripto 1.3.1, and a connection to a previous installation of MySQL 5.1.33 housed on a separate server. After the setup of the test server was complete, Omeka and Scripto required basic configuration before the preliminary transcription testing could begin.

The software configuration for both Omeka and Scripto proved to be relatively simple. There were only a handful of options within the Omeka administration console. The Omeka-specific options remained mainly untouched. However, since Scripto would dictate a large part of the user experience

for this project, the planning group did pay particular attention to Scripto's image viewer options. Scripto offers three means of presenting visual content to end users: OpenLayers, Zoom.it, and the Google Docs Viewer. Each required exploration to determine which option would work best, both for end-users as well as the behind-the-scenes workflow. Three different file types (TIFF, JPG, and PDF) were tested in each viewer, and each viewer had its strengths and weaknesses. Google Image Viewer worked well only with PDFs. OpenLayers worked with JPGs, while also being a highly customizable open source viewer. Finally, Zoom.it, a proprietary Microsoft product, worked extremely well with both TIFFs and JPGs but did not provide customizable options. While Zoom.it did not have the customizability of OpenLayers it did offer the most aesthetically pleasing viewer while also providing better clarity and a greater ability to enlarge the image. The planning group identified Zoom.it as the best choice, and JPG as the best format due to its quicker load time.

Designing the Display

Once the planning group selected the transcription tool and viewer, they began designing the public display in preparation for a pilot test. One of the benefits of the Omeka system is the adaptability of its display. The system allows different layouts and design elements for each exhibit. The planning group believed it would be easier for end users if the transcription site stood alone, separated from the Digital Exhibits. This would prevent people from getting lost in the Omeka Digital Exhibits structure. It also made it easier for programmers to create separately controlled styles and functions for the site. For the *Louisville Leader* crowdsourcing project, the Head of Web Services first set up the project as its own exhibit, or as a "simple page," before settling on creating it as a collection. The collection setting seemed to be more flexible as it allowed multiple pages to be created as needed. The project site design mimics the design used for the digital exhibits. It uses the standard colors, layout, and logo while including a distinctive project logo on all its pages.

While Omeka is set up to allow creation of exhibits through its online administrative module, the designer preferred to use the "back door" by using WinSCP to login directly to the files that needed altering. This provided direct access to the CSS and HTML codes set up specifically for the *Leader* project. In the end, the designer created two page templates: one for listing the articles available to be transcribed, and a template for displaying the article and transcription box.

With issue-level metadata already available in CONTENTdm, and the goal of the transcription process being keyword-searchable text, the group did not want to put a lot of additional effort into creating article-level metadata for the transcription project interface. During development, different configurations were tested with two main goals in mind: 1) provide the public with enough information about the article to choose items of interest to them, and 2) provide the Digital Initiatives Librarian enough information to know which CONTENTdm record should receive the transcribed text. In the end, the public transcription display included just the article's title and date.

Once the basic design had been created, the technology and design librarians worked collaboratively to adjust the pages using PHP, HTML, CSS, and javascript. The default design of the list page included images of the article, but the planning group decided to eliminate these. Because of the cropped nature of the articles the display was unattractive and did not provide any additional advantage to the viewer. The group also decided to have only ten items displayed at a time on the transcription page, as load time was a consideration. Another decision involved the display order: To keep the content looking fresh, the group resolved to have the newest items added to the system listed first. The default configuration did not provide notification of completed transcriptions, so the technology librarian added the functionality to email the completed transcription to the Digital Initiatives Librarian, with a subject line generated from the metadata provided in Omeka, including article title and filename. At this point,

the technology librarian also began writing a script to inform users if an article needed transcription or had already been transcribed, although this element was not available at the time of the pilot.

Once the server infrastructure setup was complete and the website designed, an internal audience tested the system. Staff members with varying levels of technological experience transcribed the first articles and provided feedback on their experience and expectations. This valuable input led to tweaks in the displays. First, they confirmed that transcription-status messages were a desired element. Second, they requested clarifications in the instructions for transcribing, and that they be made available on both the list and transcription pages. Rather than static HTML, the instructions were created as a server-side include that could be edited in a single file that would update both pages simultaneously.

[Figure 1. Articles available for transcription (Louisville Leader Transcription Project).]

[Figure 2. An article transcription page (Ibid).]

During the in-house testing and the initial launch, the transcription page included a Zoom.it display box with the article, an empty text box for typing, and then a third section that would appear when the transcriber hit the Save button. It was intended to provide the transcriber with an easy way to review the material before submitting. However, it appeared so far away from the original article that the Send button was pushed off the screen, making it harder to find. Post-launch, the group decided to eliminate the extra box that appeared upon saving when they discovered that some users were completing the transcription and saving it, but not submitting it.

Article Processing

In addition to establishing the technical infrastructure to enable crowdsourced transcription of the newspapers, policy decisions about what to have transcribed needed to be made. Volunteer transcribers

could not be expected to commit to transcribing an entire page of a newspaper at a time, so the planning group decided that articles would be a better fit with volunteers' interest and attention span.

A closer look at the articles within the *Leader* revealed that it featured some national news stories, some with attribution to news agencies and some without attribution. In order to avoid potential rights pitfalls and, more importantly, to harness the energies of the crowd toward making the unique, locally-created content more accessible, local stories were prioritized.

The graduate student intern and two student assistants (paid through the Federal Work-Study Program) were trained to select and manually segment local articles using Photoshop. This workflow involves opening a JPEG of a newspaper page, cropping an article, and saving it with a filename corresponding to issue date, page, and column (e.g., 19430306_1_4). The articles can then be uploaded into Omeka with minimal metadata of article title and filename. More than four months into the project, we have nearly 800 articles available to be transcribed at all times and several hundred articles selected but not yet uploaded. Because some student workers have been more thorough than others, some issues are better represented in the lists of articles available for transcription. About half of the newspaper issues have not yet been processed for Omeka. Given the amount of time that has already been devoted to this project, past newspaper issues are unlikely to be revisited.

Custom Omeka scripting translates the filename to a human-readable date, so transcribers can learn the date of the article and thus place it in context. If an article spans multiple pages, the additional article segments are attached to the same record in Omeka, although transcribers must save and submit each segment separately and click on links labeled "Next page" and "Previous page" to navigate between the files.

Once a transcription is received via email, the filename included in the subject line can be used to locate the issue and page of the article in the CONTENTdm Administration Module. The transcribed text

within the email is copied and pasted into the "Full Text" field for that page. After saving the page-level metadata and indexing the CONTENTdm collection, the transcribed text becomes fully searchable in the CONTENTdm collection. The transcribed article can then be deleted from Omeka.

Launch and Marketing

In addition to the technical preparations, the project required a publicity plan, so the "crowd" could learn about the opportunity to participate. In preparation for the launch, the Director of Archives and Special Collections met with her liaison in the University of Louisville's Office of Communication and Marketing. They discussed methods of marketing the project, including creating a press release and an article for the university's daily news vehicle, *U of L Today*. During this preparation the project group decided that the CONTENTdm collection homepage provided the best access for the public, as it described the historical relevance of the newspaper as well as linking to both the permanent home of the finished project and the list of articles to transcribe. Links within the CONTENTdm system can be fairly long (e.g. http://digital.library.louisville.edu/cdm/landingpage/collection/leader/), which do not present a problem while linking from online sources, but look unwieldy in print sources. Using the university's bit.ly account, the web designer created a shorter URL, http://uofl.me/lib-LouisvilleLeader, which would automatically redirect to the collection page and required visitors to type in about half as many characters. As an added bonus the bit.ly URL tracks usage so statistics could be garnered on how often visitors came in through that route.

[Figure 3. Collection homepage in CONTENTdm (*Louisville Leader Collection*).]

On February 12, 2013 the article in *U of L Today* and a post to the University of Louisville Libraries blog (http://uofllibraries.wordpress.com/2013/02/12/3888/) were published and press releases went out to local media. The following day, posts were made to the libraries' Facebook and Twitter accounts. The Director of Archives and Special Collections was interviewed by a local radio station on the day of

the launch. Those days produced the first spike in the transcription statistics, with 12 and 25 article sections transcribed, respectively. On February 19 a local television station ran a piece on the project along with an accompanying web page story. That day and the day after yielded 68 transcribed sections.

Knowing that this would be a long-term project, the marketing strategy needed to be long-term as well. While the first month produced the most coverage in local media venues, the project planners identified a variety of potential interest groups and staggered announcements to these groups. This approach proved successful as the numbers of transcriptions completed, averaging about 16 per day for the remainder of February, tended to jump in conjunction with new releases , leaping up to 49 on March 1 when the next big marketing push included an email to the university's History Department listserv and a post in *Consuming Louisville*, a local interest blog.

The marketing efforts continue, as to the results. Announcements have been sent out to blogs, community groups, professional organizations, and academic departments. In addition to the local media coverage noted above, the project was picked up by the *Journal of Blacks in Higher Education* on February 22. As of June 4, just under 4 months into the project, transcribers have submitted 1,648 article segments at an average of 14.5 transcriptions per day.

Lessons Learned

Embarking on this project has meant having faith in the public. From the start the planning group recognized the benefit of not over-managing the project. By not requiring a login, they avoided the creation and maintenance of a database of transcriber information and eliminated an extra step, and therefore barrier, in the participation process. This trust has extended to the receipt of transcriptions as well. Other than verifying that the transcribed article appears to be complete, no editing is done on the transcriptions. The newspaper was full of typesetting and editing errors, so instructions advise transcribers to maintain the integrity of the original, typos and all. In addition, reviewing each

transcription against the original would be nearly as time-consuming as doing the transcription in-house. Whatever the imperfections of the final product, the transcribers' work is making the content more accessible than it would have been without their contributions.

This faith in the public has not gone unrewarded. Thus far, only one of the entries into the transcription site has not been a reasonably accurate transcription. One user added commentary on the current state of a public housing project that was new (and full of hope) in a 1940 issue of the *Leader*. It seems that the commenter thought the transcription field was a comment field like those found on online editions of contemporary newspapers.

The problems that did occur had little to do with the integrity of the crowd. The tricky issues revolved around how people used the system, particularly the order of the materials transcribed. The planning group anticipated that users would select articles based on the headlines, looking for content that would be interesting; however, it seems that many users click on the first available article. Since recently added items appear at the beginning of the list of articles available for transcription, articles added first tend to stay in the system longer. This is especially obvious after a weekend when the articles just added in the preceding week have a "Transcription in Progress" status while articles that have been in the system for months remain untouched. While the general trend seems to be to transcribe the first available article, some "power users" who visit the site regularly do consciously seek out articles from the same issue to feel a sense of context and completion.

Another challenge has been articles spanning multiple pages. While attaching multiple article segments to the same Omeka record simplifies the workflow for library staff, it seems that many transcribers either do not notice the "Next page" link within Omeka or feel a strong enough sense of accomplishment from transcribing one part of the article that they do not feel compelled to follow the link to complete the second part of the article. Once a transcriber begins an article, however, the red

"Transcription in Progress" message steers other transcribers away from that record, with the result that the second page stays in Omeka. Frequently, the Digital Initiatives Librarian completes them in order to move them out of the queue. The planning team continues to look for ways to improve the user's experience in working with multi-page articles.

Another issue discussed by the planning team was whether crowdsourcing, or at least this approach to crowdsourcing, actually reduces the amount of staff time required from the host institution. The manual work to process transcriptions and to upload new articles initially required all available working hours of the Digital Initiatives Librarian and three students, and continues to be part of the daily routine of the Digital Initiatives Librarian. As an experiment, one student worker took newspaper issues and transcribed them directly, rather than cropping the images with Photoshop, loading the articles into Omeka, and having members of the public transcribe them. While she was only able to get through 24 newspaper pages in about 34 hours, the direct transcription approach proved efficient in other respects. The text files into which she typed her transcriptions, page by page, were easier to copy and paste into CONTENTdm than the transcribed sections. Since she was transcribing whole pages, this made loading the articles into CONTENTdm more efficient as the pages needed editing only once. Items that come in from public transcribers often require editing the same page multiple times.

While this project involved more work for staff members and students than initially anticipated, there were many benefits. First, it provided the ULL, and Archives and Special Collections in particular, with positive publicity. While the immediate objective of the publicity surrounding the project was to promote participation, it had the side effect of advertising the work of the unit and its commitment to providing resources of interest and use to the community.

The project has also allowed a wide range of individuals to engage with material held by ASC – materials that they would likely not have worked with otherwise. Users have included students at UofL,

professors in other states, librarians in other countries, colleagues, and local retirees. One particularly dedicated, detail-oriented, and thoughtful participant noted,

> I am enjoying this… I know I am making a contribution, and in the process I am getting a good look at history from a different perspective. Because I have generally transcribed in a consecutive timeline, I feel that I have known some of these people, their clubs and church work, etc., as well as some of the issues that had meaning for them.

The project also enhanced access to a resource that researchers find useful. As one blog commenter put it, "I am excited about this endeavor! I am currently doing research on Louisville and have had a hard time obtaining what I am looking for. I feel this will be a great gateway to my research! Thank you so much for doing this!"

Future Considerations

The planning group envisions future enhancements, such as more granular information about the completeness of a transcription when page one of an article has been transcribed but page two remains. An improved public display could list both sections of a multi-page article. Additionally, the development of a means to feed transcriptions directly from Scripto to CONTENTdm would be welcome. This would reduce the amount of intervention by staff members; the lack of spam or grievously inaccurate transcriptions indicates that this would not significantly reduce the quality of the end product.

Conclusion

Despite challenges that remain, the planning group views the *Louisville Leader* project as a success. With no additional funding, the project created a great deal of goodwill with the community, giving the public a meaningful way to interact with and improve access to an important collection; giving researchers improved access to unique primary source materials, and giving library staff a sense of

excitement and pride for their involvement in such an interesting and fulfilling project. It has also provided the public with enhanced access to an important resource for researchers, whether they are studying social history or their family tree. The planning group is satisfied with the work completed in this phase of the project, and looks forward to seeing how other institutions build on this work, or find other paths to the same end.

References

Brabham, D.C. (2013). *Crowdsourcing.* Cambridge, MA: MIT Press.

Causer, T., Tonra, J. & Wallace, V. (2012). Transcription maximized; expense minimized? Crowdsourcing and editing *The Collected Works of Jeremy Bentham*. *Literary and Linguistic Computing, 27(2)*, 119-137.

Center for Bibliographic Studies and Research, University of California, Riverside (2008). *California Digital Newspaper Collection*. Retrieved from http://cdnc.ucr.edu/cdnc

Citizen Science Alliance (2009). *Galaxy Zoo 2*. Retrieved from http://zoo2.galaxyzoo.org/

Holley, R. (2009). How good can it get?: Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine, 15(3/4).* doi:10.1045/march2009-holley

Howard, J. (2011, May 23). Gaming the archives. *Chronicle of Higher Education.* Retrieved from http://chronicle.com/blogs/wiredcampus/gaming-the-archives/31435

Howe, J. (2006). The rise of crowdsourcing. *Wired magazine*, *14 (6)*. Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

Howe, J. (2008). *Crowdsourcing: Why the power of the crowd is driving the future of business.* New York, NY: Crown Business.

Lang, A.S.I.D. & Rio-Ross, J. (2011, October 31). Using Amazon Mechanical Turk to transcribe historical handwritten documents. *Code4Lib Journal, 15*. Retrieved from http://journal.code4lib.org/articles/6004

Louisville Leader Collection. (2013, June 27). Retrieved from http://digital.library.louisville.edu/cdm/landingpage/collection/leader/.

Louisville Leader Transcription Project. (2013, June 27). Retrieved from http://exhibits.library.louisville.edu/omeka/items/browse/?collection=10.

National Archives and Records Administration (n.d.). *Pilot Project*. Retrieved from http://transcribe.archives.gov/

National Institute of Standards and Technology (n.d.). *NIST Digital Archives: Digital collections of the National Institute of Standards and Technology*. Retrieved from http://nistdigitalarchives.contentdm.oclc.org/

National Library of Australia (n.d.). *Trove*. Retrieved from http://trove.nla.gov.au/

New York Public Library (n.d.). *What's on the menu?* Retrieved from http://menus.nypl.org/

Saxton, G.D., Oh, O. & Kishore, R. (2013). Rules of crowdsourcing: Models, issues and systems of control. *Information Systems Management, 30* (*1)*, 2-20. doi: 10.1080/10580530.2013.739883

Roy Rosenzweig Center for History and New Media, George Mason University (2013). About. *Scripto: A community transcription tool*. Retrieved from http://scripto.org/about

Saylor, N., & Wolfe, J. (2011). Experimenting with strategies for crowdsourcing manuscript transcription. *Research Library Issues, December 2011*, 9-14.

Tanner, S., Munoz, T., & Ros, P.H. (2009). Measuring mass text digitization quality and usefulness. *D-Lib Magazine, 15 (7/8).* doi:10.1045/july2009-munoz

University College London (n.d.). *Transcribe Bentham*. Retrieved from http://blogs.ucl.ac.uk/transcribe-bentham/

University of Iowa (n.d.). *DIY history*. Retrieved from http://diyhistory.lib.uiowa.edu/

Zarndt, R. & Geiger, B. (2013, April). *Productivity of the crowd*. Presented at the American College and Research Library conference, Indianapolis, IN. Retrieved from http://www.slideshare.net/cowboyMontana/productivity-of-the-crowd-slides-acrl-20130412-18642176

Zooniverse (2012). *Old weather: Our weather's past, the climate's future.* Retrieved from http://www.oldweather.org/

Figures