

Crowdsourcing Documentary Transcription: an Open Source Tool

I. Project Description

Historians and scholars rely on documentary editing projects to provide necessary access to primary sources, but realistically we cannot expect unlimited funding to support the armies of paid editors to carefully transcribe extensive archives indefinitely.ⁱ (See *Appendix A: Endnotes*.) Many efforts to publish the historical documents of the nation's founders have been decades in the making and still have significant volumes yet unpublished. For example, the *Papers of George Washington* project began in 1969 and has published 52 of 90 expected volumes in the past 40 years. On a smaller scale, the *Papers of James Madison project*, which represents an archive of roughly 27,000 document, began in 1956 and has not published one third of the expected volumes (15 of the 45).ⁱⁱ This important work has taken hundreds of thousands of hours of dedicated staff time and millions of dollars in funding. Proceeding chronologically, documentary editing projects may face financial hardships before they transcribe and publish the heart of the historically significant elements in a collection.

Thus, we are left with the problem of how to make tremendously important historical collections, particularly those comprised of handwritten documents, more accessible to researchers. Though transcription is but one element of the documentary editing process, it is an essential one, and one that is ripe for innovation. Digital archives and documentary projects need a viable solution that lowers both the cost and the investment of staff time involved with transcribing large numbers of historical documents.

Therefore, the Center for History and New Media (CHNM) at George Mason University (GMU) seeks a Level II Digital Humanities Start-up Grant of \$49,215 to build an open source transcription tool that will allow all users easily to submit transcriptions and knowledge to digital archives. The resultant tool will draw upon the wisdom of the thousands of interested researchers, scholars, and students who work with digital archival materials everyday. Gradually, their combined work will enhance the discoverability and usefulness of the archives without significantly adding to project costs.

As our test case, we will implement the transcription tool with our digital editorial project, the *Papers of the War Department, 1784-1800 (PWD)* <wardepartmenpapers.org>. On the night of November 8, 1800, fire devastated the United States War Office, consuming the papers, records, and books stored there. Two weeks later, Secretary of War Samuel Dexter lamented in a letter that "All the papers in my office [have] been destroyed." From the perspective of historians, the loss was monumental since in many respects the documents lost in the fire constituted the first "national archive" of the United States. PWD is an ongoing digital editorial project that encourages historical scholarship of this lost period in the Early Republic by restoring the archive online and making the entire collection accessible and fully-searchable to a wide audience of scholars, students, teachers, and the general public. This website officially debuted in June 2008 with nearly 45,000 document images and metadata, and the site is on target to receive 49,000 unique visitors in 2009.

A scholar searching the PWD archive for documents that include the phrase "Indian," "treaty," "negotiation," "boundary," and other terms will easily find a May 2, 1788 letter from Arthur St. Clair to Charles Thomson discussing negotiations with Indian tribes. Since the PWD archive contains only keyword metadata, rather than full transcripts of the documents, the engine cannot rank the search results according to relevance and significance. A visitor performing that keyword search will likely receive more than 150 items along with the St. Clair-Thomson letter. Many of those documents are brief, mentioning the terms only once, and are therefore likely to be of less interest to researchers. Without a full transcription, the search engine cannot distinguish between these short items and the far more relevant letter, which mentions each term dozens of times over its five pages.

An Opportunity to Capitalize on the Wisdom of Users

While the staff at the PWD will never have the capacity to transcribe a significant number of the documents in the archive, our yearly site traffic suggests that we have a vast resource of scholars and researchers to draw upon to help us with this task. Everyday researchers examine documents in the archive,

often making rough transcriptions to use in their own work. We routinely get emails from researchers suggesting improvements to the archive's metadata based on their work with the materials.

In proposing to open up the transcription process to crowdsourcing, CHNM is drawing upon the example and success of a host of other successful ventures. Though the goals and needs of digital archive and documentary editing projects are distinct from the goals of these crowdsourcing ventures, they do offer promising glimpses of the efficiencies and outcomes for our purposes. Each taps into the interests and passions of a segment of the public to contribute to the accumulated value and content of a project—whether software development or knowledge aggregation.

Wikipedia, the free online encyclopedia that is written and maintained by users around the world, is by far the most well-known instance of successful crowdsourcing. The site currently receives roughly 65 million visitors each month and has more than 75,000 active contributors who create and edit over 13 million articles—more than 3 million of which are in English.ⁱⁱⁱ From its founding, teachers, parents and scholars have worried about the accuracy and content of Wikipedia, especially since articles from the free encyclopedia have long been the first result to appear in most search engines. Yet, the source of the occasional errors and incoherence of Wikipedia is also its power—the tremendously fast-moving contributions and collaboration of users means that errors are unlikely to remain for long before they are corrected.^{iv}

Flickr Commons provides another example of tapping the collective knowledge of interested members of the public. Beginning with a seed contribution by the Library of Congress in January 2008, cultural institutions from around the world have contributed images and associated metadata to the Flickr Commons collection. Collections of images are freely available to the public, who can tag and comment on each image. A year ago, the Library of Congress assessed their participation in the pilot project, reporting that over 2,500 Flickr users had left more than 7,000 comments on almost 3,000 images. Library staff culled important corrections and additions to captions, titles and the identification of individuals that were then incorporated back into the metadata of more than 500 items by August 2008. Based on these positive interactions with the public, the staff advocated continuing to draw upon public knowledge through Web 2.0 interactive projects, declaring: “The benefits appear to far outweigh the costs and risks.” The Library of Congress's success helped to encourage the Smithsonian Institution to join Flickr Commons, as a result, in the period between June and December 2008, their photographic contributions received over 625,000 views with many user comments and tags. Thus, trusted cultural heritage organizations are beginning to recognize the powerful ways that interested members of the public, scholars, and educators can contribute to the knowledge base related to collections.^v

To bring this type of crowdsourcing to the world of digital archives, CHNM will create a transcription tool that makes it easy for users to contribute transcription text for the documents they read and use in their research. Locating a record in the digital archive, a researcher currently has access to the metadata for the letter and a link to a high resolution image of each page of the document. With the addition of the transcription tool, a user will also have the option to contribute a transcription of the document, to engage in discussion of the item, and to review the editing history of the transcription, if there is one.

Should one of these users click on the link to contribute a transcription, she will be brought to a new webpage consisting of two panes. On the left side of the page, a document viewer will provide access to the high resolution images of the letter, with the capacity to zoom in on the images and page back and forth between them if there are several. The right pane of the screen will include a login and account interface for first time users, which will protect the archive from spam and vandalism. Once the user has logged in, she will be presented with a simple wiki editing window into which she can enter her transcription of the document, drawing on the formatting conventions available through MediaWiki. Once she saves her transcription, the most recent version will appear with the letter's metadata on the main record page. If the user has significant knowledge about the St. Clair–Thomson letter, upon returning to the main item record, she could click on the discussion link to leave some notes about the document. Similarly, if she was unsure about some element of her transcription, she could note that for future users. (*See Appendix B: Mock-ups.*)

Unlike traditional crowdsourcing ventures, opening transcription of digital archives will necessarily draw upon the contributions of a specialized audience of users doing different types of history. The user base for the PWD is composed primarily of scholars interested in the Early Republic, military, and Native American history—advanced undergraduate students, graduate students, junior scholars, and senior scholars,

all of whom have their own reasons for creating transcriptions of particular documents. Certainly there are also members of the interested general public who use the archive in their work to research family history and other related topics. The addition of the transcription tool to the PWD digital archive will allow us to reach out to an entirely new audience that will include public historians in training. The open interface will allow documentary editing students, in particular, to operate on a live archive, practicing the techniques and elements of faithful archival transcription that they are learning in their classes and seminars. As other digital archives begin to implement the tool, the array of document types and conditions on which proto-editors can practice will grow. Together these users compose a specialized audience who can contribute both their transcriptions and insights to the collective knowledge base of the archive.

Public transcription of digital archival material will not immediately produce perfect results. However, just as the errors in Wikipedia are quickly corrected by other users, the open access provided by the transcription tool will allow the editors and other users to correct mistakes as they become apparent. The editors will provide oversight, reviewing document discussions, mining research notes to add to existing metadata and cross referencing.

This scenario suggests that there will be significant benefits for interested users, whether they are scholarly researchers, students and teachers, or members of the interested public. First, general users will immediately benefit from the improved search capacity that full-text transcriptions will yield. Second, students of documentary editing and their instructors will have access to a space where they can participate in the process of transcription while dealing with actual archival sources. Students will have the experience of approaching the documents as editors and contributing to the work of the project. Third, the modified wiki interface will allow users to engage in an online conversation about particular documents. Researchers will be able to leave notes for the editors, both about the document transcription, but also about the item's metadata and its relationship to other items in the archive. Thus, the public transcription tool will open up a venue to encourage scholarly communication, collaboration, and community building based on the documents.

Similarly, crowdsourcing transcription for digital archives will provide significant benefits to documentary editors and to the broader field of public history. Archival staff will be able to draw on the publicly contributed transcriptions to form a base for their editorial work. Editors may start with a rough transcription provided by interested users, and then apply the techniques and expertise of their training to produce a corrected transcription quickly. Moreover, allowing the public to contribute document transcriptions to digital archival collections will have real-time benefits for the accessibility of the documents. Each transcription contributed to the archive will immediately be available to the collection management system's search engine. The result will be ever-improving discoverability. The full text of the documents will allow the search engines to surface the most relevant documents by better weighting their results. Also, crowdsourcing transcriptions will allow editors to better understand the ways in which users interact with their archive. Necessarily the documents that are of the most interest to users will be transcribed most quickly and fully. Thus, the public contributions can serve as a barometer of the most interesting materials within a particular collection.

In sum, crowdsourcing transcription has the possibility of changing what is possible in the world of digital documentary editions by facilitating faster, more complete, useful work drawing on fewer resources.

II. Environmental Scan

The call for this type work has been issued from within the archival community. For instance, "The Interactive Archivist: Case Studies in Utilizing Web 2.0 to Improve the Archival Experience" published by the Society of American Archivists, provides a useful summary of the interests of the archivists in social networking and the usefulness of tagging, commenting/reviewing and rating services, but does not mention strategies for incorporating users into the archival process at the transcription or description level that remains the domain of archivists. Similarly, in his 2007 article, "Archives of the People, by the People for the People," Max J. Evans points to "the concept of commons-based peer-production as a means of turning collections inside out. It encourages archival institutions to reinvent themselves, and, in collaboration with other archives and with other types of organizations, to organize archival work in concert with a curious and interested public." Though inviting the public to participate in the transcription process of colonial-era manuscript collections is unprecedented, other projects have successfully employed crowdsourcing online.

The most prominent examples are connected to the USGenWeb Project. Starting in the late 1990s, the USGenWeb Archives Project was developed to publish actual transcriptions of public domain records on the Internet. This massive undertaking is a result of cooperative efforts among volunteers who share electronically files of census records, marriage bonds, wills, and other public documents, while others transcribe these documents.”^{vi} These studies and projects demonstrate that the archival community is primed and ready to move into the world of crowdsourcing, but they simply lack the tools.

III. Project History

CHNM has been the home of the PWD project since 2006. In the course of that time, the document processing and editorial work has been supported by grants from NEH and the National Historical Preservation and Records Commission. We are committed to the long-term sustainability of the archive and expect the editorial work to continue for at least another three years. Additionally CHNM, has over a decade of experience with creating and maintaining digital archives, such as large-scale, preservation-grade digital archives, such as the Sloan Foundation funded *September 11 Digital Archive* and the *Hurricane Digital Memory Bank*, and the NEH funded *Bracero History Archive*, which have fostered a public and interactive model for documenting and preserving the past. This experience has led us to understand the need for archival transcription, and the vast possibilities of crowdsourcing to begin to meet those needs. The work supported in this grant will contribute to these and subsequent efforts at building digital archives, both those projects managed by CHNM and those ventures undertaken by the rest of the digital humanities community.

IV. Plan of work

Our work on the crowdsourcing transcription tool will span one year and will be composed of six phases. Each phase will involve a significant commitment of staff time, as indicated in the budget. (*See also, Budget Narrative*):

- **July 2010-August 2010: Planning:** During the first three months of the grant, we will do intensive planning work using the PWD website as a case study. First, the project team will create a set of wire frames and user interface scenarios that will help predict the best workflow for moving users from the existing digital archive to the transcription interface. Also, we will establish the best methods for using MediaWiki as the database and formatting engine behind the transcription tool. This will require a thorough knowledge of the MediaWiki application programming interface (API). The result of this phase of work will be a set of workflow diagrams and schema that will guide the production of the transcription tool, which will allow it to function with these common collection management systems.
- **August 2010-January 2011: Production of the transcription tool:** After the initial planning stage the web development team will embark upon building the transcription tool for the PWD website. This will represent the bulk of the technical work for the project. Creating the tool will require two steps (*See Appendix B: Mockups*):
 - We will build a generalizable tool for document transcription, comprising a Web interface powered by a lightweight PHP library designed to interface the MediaWiki API and any compliant content management system (CMS).
 - The **Web interface** will have two major components: a document viewer and an editing window for the transcript input. The **JavaScript document viewer** will make it possible for users to page through all of the images that accompany a multi-page document, and to zoom in on high resolution image so that they can better decipher difficult handwriting and marginalia. The **transcription form** will include a login interface, a editing window, and a link to helper screen that summarizes the text formatting conventions.
 - The **PHP library** will have two major components. The MediaWiki API client will communicate with the MediaWiki API server. The CMS adapter interface will define the requisite data needed from the CMS for the transcription tool to function properly.

- We will integrate the transcription tool into the PWD archive by writing an **adapter** for the tool and an **API server for PWD**, so that we will be able to access the high resolution images for each document via an XML or JSON feed. We will write a script that uploads the most current version of the transcription into the PWD database, allowing us to publish and search the most up-to-date version of each document transcription. Other developers will be able to write similar adapters tailored to their specific CMS (e.g. Omeka, Drupal, Wordpress, Joomla, custom systems).
- **January 2011-June 2011: Implementation:** Once we have built the elements of the transcription tool, we will implement it with PWD and invite volunteers to test a beta version of the tool. In particular, we will invite the many library, museum and archive users of our *Omeka* web publishing software. We will provide our testers with two levels of support as they use the transcription tool with their collections:
 - We will set up a developers forum so that test users can raise questions and concerns about the tool as they are implementing it with their systems.
 - Our web development team will hold “office hours” in IRC Chat on Fridays during the implementation and testing phase of the project, so that test developers can have ongoing access to advice and help as they work to install and customize the tool for their digital archives.
- **March 2011-June 2011: Evaluation:** We will also conduct several layers of qualitative evaluation of the transcription tool. For our volunteer testers, we will create a survey that covers several areas of inquiry, including the quality of the experience with the user interface for editors, the function and stability of the tool’s code and features. This qualitative information will combine with quantitative information gathered from server logs on page visits and the cumulative total of transcription edits, user accounts created, and documents with discussion to form a picture of the potential for crowdsourcing digital archival transcription. After gathering this feedback from our testers and users, we will make any necessary changes in the functionality of the tool before we do our official release and publicity for the project.
- **March 2011-June 2011: Authoring the *Guide to Crowdsourcing Documentary Transcription*:** Upon completing the construction of the elements of the transcription tool and implementing it with the PWD, the project team will embark upon writing a guide for other editors who wish to implement the transcription tool with a digital archive. We will provide explicit documentation on implementing and customizing the tool for use with a variety of collection management systems, including tips on creating adapter plugins for various platforms, skinning the tool to match a website design, and modifying the user account login and permissions structures. On the practical side, we will provide advice on encouraging contributions from the research community, and the demands managing the oversight of transcriptions and document discussion notes by digital archive editors.
- **May 2011-June 2011: Dissemination:** After implementing any necessary revisions to the transcription tool, we will make an official release of the tool and the *Guide*. (*See Dissemination below for a full description of our distribution and publicity plan.*)

V. Staff

CHNM’s extensive experience in digital projects equips it to manage the planning, development, and dissemination of this public transcription tool. The project builds directly on three core strengths of CHNM: 1) our development of widely used open source digital tools, including *Omeka* (recent winner of a Mellon Award for Technology Collaboration) and *Zotero* (named among PC Magazine’s “Best Free Software” two years running); 2) our creation of large-scale, preservation-grade digital archives which have fostered an interactive model for documenting and preserving the past; 3) our use of the web for more than a dozen years as a vehicle for engaging broad and diverse audiences, often in partnership with libraries, archives, and museums. This long-standing track record of outstanding and innovative work in digital history projects makes us the ideal candidates to build and test an application for crowdsourcing document transcription. (*See Appendix C: Organizational Profile.*)

Key personnel (*See Biographies.*):

Sharon Leon (Primary Investigator) is Research Assistant Professor of History at George Mason University and Director of Public Projects at the Center for History and New Media. She will assist in supervising the project and will coordinate its technical aspects, devoting 10% of her time to the project.

James Safley (Technical Director) is the architect of the *Papers of the War Department's* database and the designer of its powerful and flexible search engine. He will do the work of programming the tool, creating the PHP library, and advising the technical staff of our testing partners, devoting 25% of his time to the project.

Chris Raymond (Web Designer) is an award-winning graphic and web designer. She will create and style the user interface for the tool, including the image viewer, and will devote 12% of her time to the project.

VI. Final Products and Dissemination

Products:

- An open-source transcription tool that brings together 1) a modified wiki package with versioning, user login and authentication, and the ability for discussion; 2) a JavaScript-based document viewer that allows users to zoom in on high resolution images of documents.
- A model for creating CMS adapters that will connect the transcription tool to potentially any content management system (Omeka, Drupal, Wordpress, etc.), allowing both for the call of the document images and the return of the transcription text to the main archive.
- A *Guide to Crowdsourcing Documentary Transcription* that provides both technical and practical advice for editors of digital documentary archives who wish to deploy the tool.

Dissemination:

We will widely disseminate our work both to the world of documentary editors and the world of digital humanities, through CHNM's *Labs* portal <chnm.gmu.edu/labs/>. The CHNM website receives significant traffic each year, accruing a total of over 11 million unique visitors in 2008. The products of "Crowdsourcing Documentary Transcription" will fit nicely with the other cutting edge projects housed at *CHNM Labs*, such as *Mobile for Museums* < chnm.gmu.edu/labs/mobile-for-museums/> a project funded by the Samuel H. Kress Foundation to study the use of mobile devices in the museum world and to develop several innovative prototypes to improve that usage. From *CHNM Labs*, we will link to the source code, publish the PHP library, and post the *Guide to Crowdsourcing Documentary Transcription*. Since the transcription tool is an open source project, we will publish the source code at Google Code under the open source General Public License (GPL). This license means that anyone can use the code and modify it to suit his or her needs.

CHNM sponsors a number of widely read professional blogs, and these blogs have proven to be a very effective means for rapidly disseminating news to those interested in digital tools. We will also make use of the *History News Network*, a CHNM affiliate that sends a newsletter to 13,000 scholars, educators and interested members of the general public three times per week. Finally, we will reach out to documentary editors who may wish to use the tool with their archives by publicizing it with important discussion networks (e.g. SEDIT-L), and by making presentations at the Association for Documentary Editing and the Society of American Archivists' annual meetings.

Appendix A: Endotes

-
- ⁱ Réjean Plamondon, Sargur N. Srihari, “On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, January, 2000.
- ⁱⁱ See, “Project and Volumes,” *The Papers of George Washington*, <<http://gwpapers.virginia.edu/project/index.htm>>, and “About,” *The Papers of James Madison*, <<http://www.virginia.edu/pjm/description1.htm>>.
- ⁱⁱⁱ “About,” *Wikipedia* <http://en.wikipedia.org/wiki/About_Wikipedia>.
- ^{iv} Roy Rosenzweig, “Can History be Open Source? Wikipedia and the Future of the Past,” *The Journal of American History* Volume 93, Number 1 (June, 2006): 117-46. Available at <<http://chnm.gmu.edu/essays-on-history-new-media/essays/?essayid=42>>.
- ^v Michelle Springer, et. al., *For the Common Good: The Library of Congress Flickr Pilot Project, Final Report* (October 30, 2008): 25, 30, 36. Available at <http://www.loc.gov/rr/print/flickr_report_final.pdf> and Kalfatovic, et al., “Smithsonian Team Flickr,” *Archival Science*, (October 2009): 11-12.
- ^{vi} J. Gordon Daines III and Cory L. Nimer, *The Interactive Archivist: Case Studies in Utilizing Web 2.0 to Improve the Archival Experience* (Society of American Archivists, May 18, 2009), <http://lib.byu.edu/sites/interactivearchivist/>; Max Evans, “Archives of the People, by the People, for the People,” *American Archivist* 70, no. 2 (2007): 387-400; “The USGenWeb Project - Special Projects,” <http://www.usgenweb.org/projects/index.shtml>.