

**Interim Performance Report for National Historical Publications and Records Commission
Grant NAR10-DS-50017-10
Implementing Crowdsourcing Transcription for Archives and Documentary Editions
July 31, 2011**

Executive Summary:

Since January 1, 2011, the Center for History and New Media has made significant progress on the development and implementation of *Scripto* <scripto.org>, our open source tool for crowdsourcing documentary transcription. We have had a successful implementation of the tool with the *Papers of the War Department Project* and we have released an alpha plugin for use with *Omeka*, which is currently in testing.

Ongoing Development of the Basic Tool:

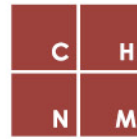
In January, Web Developer Jim Safley made the code for *Scripto* available for testing through GitHub <<https://github.com/chnm/Scripto>>. This site allowed any interested user to download the code, install it, comment on it, and contribute to development. As a result testers from Alberta, CA, Austin TX, Cambridge MA, and elsewhere downloaded the code and submitted a number of helpful development issues that eventually resulted in improvements to the functionality.

***Papers of the War Department* Implementation:**

As we approached the public launch of *Scripto* in March, Project Director, Sharon Leon, authored a blog post to explain CHNM's considerations upon opening up a project to contributions from the community. The piece articulated four goals for the project, which remain central to our work:

- 1) Text contributed by the community is essential to improving the findability of individual documents;
- 2) The community-created transcriptions will provide editors with a first pass at transcription that can later be improved if resources permit;
- 3) The selection of documents for transcription by the community provides valuable insight into their interests in the archive; and
- 4) Opening the archive to community work allows us an opportunity to create community around historical materials.

These principals continue to guide our work as we engage with community transcribers and their contributions.

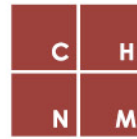


On March 17, 2011 CHNM launched the first implementation of *Scripto* with the *Papers of the War Department* project. This implementation involved a minor redesign of the *PWD* site to include access to the transcription functionality, and several new pages to support the process. With the redesign, users now have access to a sign-up page, a list of suggested documents for transcription, and a set of guidelines for their work. *PWD*'s new functionality garnered immediate attention both in the digital humanities community and in the larger technology community. For example, the project received notice in a post by Curt Hopkins entitled "Crowdsourcing the Preservation of the U.S. War Papers" <http://www.readwriteweb.com/archives/crowdsourcing_us_war_papers.php> on *Read, Write, Web*, a leading source of technology news.

The enthusiasm of *PWD* users for the transcription project has exceeded our expectations. Between the launch of the functionality and June 30, we have had 240 users sign up for transcription accounts. Those users have begun transcription of 210 documents, each of varying numbers of pages, and have substantially completed the work on 175 documents. This work is proceeding at the pace that we expected, slowly but steadily. James Halabuk, the Digital History Associate responsible for overseeing the contributions and editing the work, has engaged in 105 conversations about the documents and the process of transcription through the "Talk" feature of the interface. These interactions have resulted in improving the work of the transcribers and reassuring them that they have the support of the editorial team should they need it. Additionally, Halabuk has begun a series of blog posts for the *PWD* site updating the community on the project's progress, such as "Three Months On—A Transcriber Spotlight" <<http://wardepartmentpapers.org/blog/?p=374>>, which introduces one of our most prolific contributors, Patricia Gerard. Finally, while site usage statistics are difficult to generate for the transcription elements of *PWD*, we can confidently count at least 130,000 pulls on the *Scripto* API, and roughly 3,000 views of individual user transcription records. This suggests healthy and steady return traffic to the site from our registered users.

***Scripto* Plugin for Omeka:**

Upon completing the initial implementation of *Scripto* with *PWD*, Jim Safley turned his attention to building a *Scripto* plugin for *Omeka*, CHNM's popular web publishing platform for cultural heritage materials. Given that *Omeka* is a standards-based content management system, this plugin presents the possibility of opening community transcription functionality to a wide audience. Since the data model for *Omeka* differs significantly from that for *PWD*, Safley had to explore a number of complex possibilities in creating the plugin. While the creation, collection, and transmission of the transcription was relatively similar, the negotiation of calling the source files into the *Scripto* interface was significantly more complicated. Unlike *PWD*, *Omeka* allows for the inclusion of many different kinds of files in addition to page images. Thus, Safley decided to build in a variable option for *Scripto* to render any file type allowed by *Omeka* in the viewer window. This required the inclusion of player capacity for audio and video files, as well



as an all-purpose document viewer for pdfs and other document formats. All of these developments were available to users who were following the commit log for the software on GitHub < <https://github.com/chnm/Scripto/commits/master>>.

Satisfied with the results of our internal testing, Safley offered the user community the opportunity to test an alpha of the *Scripto for Omeka* plugin in early July. His blog post on the *Scripto* site included installation instructions, so that interested users could quickly implement the plugin. And, that's exactly what they did. Within hours of the announcement about the alpha release Miriam Posner, the Mellon Postdoctoral Fellow at the Emory University Libraries, had *Scripto* up and running on her own personal site. She offered a number of screenshots through twitter for other users to examine:

- Transcribe Page < <http://lockerz.com/s/117613621>>
- Revision Versions < <http://lockerz.com/s/117856511>>.

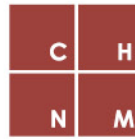
Posner's ability to install the plugin and have it running on her site the same day as the release suggests that other users should not have significant difficulties using the software. In this vein, we have notified all of the institutions that have expressed interest in testing the software that the alpha is available and that we would appreciate their thoughts on it.

Outreach:

We have approached outreach for *Scripto* in three main ways over the course of the last six months. First, we have provided periodic update through blog posts. On the *Scripto*, blog Sharon Leon and Jim Safley have made readers aware of the progress on the generalized code libraries. James Halabuk has handled all of the transcription related updates on the *PWD* blog. Second, we have established twitter accounts for both *Scripto* and *PWD*, which serve to amplify the reach of the blog posts, and to allow for interaction with followers. Finally, we have established a Google Group specifically for web developers who want to raise issues about the code and functionality. Though the group is currently quiet, we expect it to serve as a site for discussion and support as testers begin to implement and modify the software.

In addition to these general and ongoing outreach efforts, CHNM Web Developer Patrick Murray-John discussed *Scripto* with the attendees at the Maryland Institute for Technology in the Humanities' API Workshop < <http://mith.umd.edu/apiworkshop/>> in February. Workshop participants included leading programmers and developers from Google Maps, Open Library, Data Unbound, Freebase, and digital humanities centers throughout the country. Murray-John's presentation allowed the participants to consider *Scripto* in the wide array of tools at their disposal to access and enhance existing streams of digital content.

While the *PWD* implementation has garnered the majority of the focus of outside attention for *Scripto*, the tool is also beginning to gain some traction as individuals and institutions consider their own projects. Thus, *Scripto* has been mentioned in a host of blog and forum posts from a wide variety of interested parties, including the European Journalism Centre < <http://www.ejc.net/resources/featured/scripto/>>, and in the Ideas Incubator of the



Center for History and New Media
[Building a Better Yesterday, Bit by Bit]

Open Knowledge Foundation <<http://ideas.okfn.org/ideas/59/generic-transcription-tool>>.
This warm reception suggests that there is a great interest from the digital humanities community at large in crowdsourcing transcription.

Overall, we are pleased with the progress of the project and we have high hopes for the coming year. We welcome further input from and discussion with the NHPRC staff as our work continues. Please feel free to contact either of us at any time, should you wish to visit the Center or to discuss developments in digital humanities more broadly.

Sincerely,

Sharon M. Leon
Director of Public Projects and Associate Professor,
Roy Rosenzweig Center for History and New Media
sleon@gmu.edu
703-993-3831