

George Mason University, Center for History and New Media

Papers of the War Department, 1784-1800

Progress Report, November 1,2006-January 31, 2007

Over the past three months, the Papers of the War Department 1784-1800, now located at George Mason University's Center for History and New Media and led by Christopher Hamner and Roy Rosenzweig, has continued its progress in updating, extending, and preparing to make available the work begun at East Stroudsburg University. Work is progressing well on both the technical and editorial sides of the project. On the technical side, our work is focused on three facets of the project: (1) migrating ESU's digital data to a new database and verifying the integrity of the migrated data, (2) preparing an interface with which the Mason editorial staff can enter new data into the database, and (3) beginning to design a search engine that will allow researchers to access the data with maximum flexibility. On the editorial side, Mason's documentary editing staff continues to evaluate the work already completed at ESU for accuracy and to process new documents into the database in discussion with Consulting Editor Jane Turner Censer and with assistance from former editor Ted Crackel.

With the project's physical files relocated and organized in a dedicated space on Mason's Fairfax campus, work on the technical side of the project during this three-month period continued with the critical—and enormously complex—task of migrating the data to a new database sufficient to the online publication of the papers. In its ESU incarnation, the project employed a Microsoft Access database that was not suitable for making the PWD available online, as the new CHNM team intended, nor had data been

entered consistently at ESU. Jim Safley, technical director of the project, and other CHNM staff continued developing an entirely new data architecture for the project that made use of open-source and standards-compliant databases and programming. With that architecture in place, Safley and the CHNM staff began to migrate the data from the Access database to the new MySQL database. A key part of the procedure involved ensuring that the migration process had not introduced errors or bugs into the dataset. The painstaking process of verification required CHNM staffers to perform spot-checks, selecting documents at random, locating the hard copies of those documents, and checking the document data against the newly-migrated information in the dataset. Because these individual document verifications had to be performed by hand, they proved quite time-consuming; fortunately, random checks of more than a hundred documents revealed no errors, convincing Safley that the data had been successfully migrated to the new database. One advantage of the arduous process of migrating and verifying the data was that it afforded Mason staff the opportunity to examine the ESU data and identify potential problems in the dataset. Assistant editors Ben Huggins and Richard Harless assisted Safley with the verification process and performed the bulk of the individual spot-checks.

In addition to the data migration and verification, Safley and CHNM technical staff completed work on the software that allows editors to input data into the new database. The result of that labor is a web-based interface that allows project staff to perform the data entry at the core of the project; the web-based format allows editors to use pull-down menus and shortcuts in places, a small but significant improvement in both efficiency and accuracy over the old methods in place at ESU. The web-based interface

has proved robust enough to handle the demands that multiple users place on it during the data entry process.

The third area of progress on the technical side of the project has been the continuing development of a web-based search engine capable of locating documents for researchers. Design of this interface, headed by Safley, proceeded with regular input of Huggins, Harless, Hamner, and other members of the CHNM staff, who provided insight into the ways that professional researchers and interested members of the public might want to access these documents and discussed features that might facilitate their searches. A beta version of the search engine incorporating their suggestions and insights is currently under construction.

On the editorial side of the project, work continued extracting data from individual documents and entering that data into the new database. This work, performed by Harless and Huggins, under the leadership of Hamner (and in consultation with Censer) occurred within a set of conventions designed to standardize the data entered into the new database. Those conventions are now largely codified, allowing us a high degree of uniformity among the new data. The process entails creating an abstract or gist of the document, including drawing from each document the key words and phrases (especially names, groups, places, issues, and ideas) that will empower users with a near-full-text search capability.

Data Entry II, as we term this phase of the work, remains a labor-intensive task, and with documents written in so many (often difficult) hands there are no shortcuts. This is one of the project's most critical efforts, since it will allow the detailed user searches that make our product so unique. It is also a vital quality control step in the

project. Undergraduate students at ESU did much of the initial basic data entry; not surprisingly, this data contains errors. The editors doing the content data entry are responsible for checking the basic metadata that has already been completed as well as insuring that there are no multiple copies of the same document and that the data records and image files are correctly linked. Our initial estimate—based on what we had been told by the staff at ESU—was that each editor would be able to process about 2 documents per hour. We have found to this point that Data Entry II requires more attention and therefore more time per document, especially as we have discovered more long documents in the archive, and it appears that Data Entry II at ESU tended (perhaps accidentally) to focus on shorter and less complex documents. At this point, therefore, we calculate that a careful editorial assistant can process approximately 1.5 documents per hour, a finding that necessarily leads us to revise the targets in our work plan downward even as we hope that our efficiency increases with experience. Three other factors that have undercut our ability to meet our original targets: the fact that the moving and reorganization of the collection has been a more time consuming process than originally anticipated; the complex work of data migration—and data checking—from the legacy system had not been factored into our calculations; and we were not aware that the ESU estimates that we used did allow time for data checking or for the training of new staff. Still, the solid base that we have set in re-establishing the project at Mason will pay later dividends. And, despite these problems and delays, we have managed to complete Data Entry II for more than 800 documents. Moreover, the addition of Mark Phillips as a third Assistant Editor alongside Huggins and Harless will allow us to edit more documents over the remaining five months of this grant than in the previous five

months. We are funding this third Assistant Editor from other cost savings in the original budget. For example, we had originally budgeted \$3,200 for travel, but we decided in the initial work plan to postpone travel to collections to search for additional documents. This and other budget savings has allowed us to hire the additional assistant editor.

A final important early goal has been to produce a more systematic assessment of the state of the project—an effort that we have begun but not completed. Former editor Crackel, with whom we have spoken regularly on the phone as well as met with in person, has generously provided his uncompensated help in this process. (Our advisory board has also provided crucial historical and technical advice on an individual basis.) This phase of the work will result in an internal report on the state of the project.

--Christopher Hamner and Roy Rosenzweig