

George Mason University, Center for History and New Media
Papers of the War Department, 1784-1800

Progress Report to NHPRC, January 1-October 31, 2006

Over the past eleven 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 made great progress in restarting and moving forward the work begun at East Stroudsburg University in the 1990s. Work thus far has ranged from physically moving the project to Mason's Fairfax campus to migrating ESU's digital data to a new database to beginning the process of editing individual documents.

We began our work in intensive discussions with former editor Ted Crackel about the state of the project at ESU and its structure. This was the beginning of the long-term process of gaining intellectual control over the project and making the transition to a new home and a new technical and editorial procedure. Consulting Editor, Jane Turner Censer, who has a decade of experience as a documentary editor, assisted Hamner and the rest of the staff in beginning to establish new editorial procedures.

One of the next key tasks was planning for the move of and then moving the documents, files, CD-ROMs, and CPUs that comprise the project from the East Stroudsburg campus in Pennsylvania to the Papers' new home at George Mason University's Fairfax campus in Virginia. Preserving the organization of the files was critical to reconstructing the project at Mason, so a team of four Mason staff members journeyed to Pennsylvania twice to move the computers and files themselves rather than contract with a private moving company.

Moving the files involved transporting more than 200 boxes of files (the hard copies of the documents obtained by ESU staff) as well as hundreds of CD-ROMs containing backups of image files, several CPUs that stored the original database, and assorted files, manuals, and correspondence that ESU staff had amassed over their decade with the project. Unfortunately, we had to at first store the files in temporary quarters at Mason because the new home for the Papers, Mason's new Research I building, was delayed in construction. But ultimately the team undertook the painstaking work of unpacking the files and carefully organizing them. The process of unpacking these files provided Mason staff with an opportunity to begin learning more about the actual content of the files and the nature of ESU's organization and labeling system, a process that continued for several weeks.

Another key—and very complex task—was beginning to migrate the data to a new database sufficient to the online publication of the papers. For the past several years, the project had used a Microsoft Access database that was not suitable for making the PWD available online (as the new CHNM team intended). In addition, data had not always been entered consistently at ESU. Jim Safley, technical director of the project, and other CHNM staff began to develop an entirely new data architecture for the project that would make use of open-source and standards-compliant databases and programming and to migrate the data from the Access database to the new MySQL database. This painstaking process has required significant manual checking and correction to ensure that the data transfer occurred without error, in addition to some digital archeology and considerable work on image processing (especially to present multi-page

documents online). This work continues at a good pace, and will be completed before the end of the current grant from NHPRC on July 1, 2007. Though time-consuming, the process of migrating the data and verifying that data has also afforded Mason staff the opportunity to examine the ESU data and identify potential problems in the dataset. Two assistant editors—Ben Huggins and Richard Harless—assisted Safley with the verification process.

As we moved ahead on the technical infrastructure for the project, we also began our first efforts in document editing and data entry. That process includes two steps: creating an interface that allows editors to input data into the new database, and developing a set of conventions governing the data to be extracted from individual documents. Huggins and Harless, under the leadership of Hamner (and with advice from Censer) have begun the time-consuming process of examining individual documents and entering relevant data into the new MySQL database. This 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 call this phase of the work, is 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 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. The project's experience so far is that error correction is a very time-consuming process. Nevertheless, we have already completed Data Entry II for 450 documents.

Another 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.) When we have completed this phase of the work, we will produce an internal report on the state of the project.

--Christopher Hamner and Roy Rosenzweig