

Work Plan. The original work plan submitted in July 2005 assumed a total budget (including cost sharing and support from the Center for Military History) of approximately \$576,000. The actual award that we will now receive is approximately \$272,000 (including cost sharing). As a result, we must naturally reduce the scope of the work plan. Nevertheless, we still believe that we can accomplish a great deal in the grant period.

The project will build on a strong base established by ESU. They have accomplished a great deal with modest funding over the past decade. Most notable of those accomplishments has been the remarkably successful effort to locate the lost papers of the War Department in dozens of collections across the country. Although the departure of Crackel in August 2004 has slowed work on the project in the past two years, ESU has still made considerable progress. Of the approximately 50,000 documents located, approximately 39,000 have been both scanned into digital form and have had basic bibliographic metadata (Data Entry I) created within the database. Much more extensive descriptive metadata (Data Entry II) has been added for approximately 12 percent of the collection (6,000 documents).

The GMU team has already completed the first step: the transfer of the project to the new location. Paper documents and files, as well as the digital files and database have already been moved into an office on the Fairfax campus. We will soon shelve and organize those files in our new offices. The GMU team has also begun the even more important transfer of intellectual and editorial control over the project. Crackel has generously provided his uncompensated help in this process, furnishing a detailed assessment of the overall state of the project, including a review of the quality of the

digital files and the metadata that has already been created and entered. When we have completed this phase of the work, we will produce an internal report on the state of the project.

The next two phases of the project—technical and editorial—will be carried out in tandem rather than in sequence. The technical work on the project will start with building a new database architecture for the project. The project files currently reside in a Microsoft Access database, which has served the project well so far, but which is not robust enough to provide public access over the web. CHNM has considerable experience with online databases and will build an infrastructure using the open source MySQL database. Then, we will transfer the records from the existing database to the new one. Because the original structure was a “flat” database, this process requires extensive reworking of the tables to create a fully relational database and extensive “normalization” of the data. In effect, we need to create an entirely new database structure so as to make the documents available on line and to finish the editorial work on the project. There are significant legacy problems in the Access database that will require a substantial investment of time to migrate the data successfully. Finally, we will write the middleware (using the open source programming language, PHP) focusing initially on creating an administrative and editorial interface to allow editors to work on the documents directly online. When that is completed, we will shift to creating the interface and programming for serving the papers to the web. The goal would be to have a working beta of the site and the database infrastructure by the end of the grant period.

In this technical side of the project, we will be starting from scratch since that phase of the project had not begun at ESU. In the editorial work, however, we will build

directly on the work already accomplished there, following the template set out by them. We will focus on two main tasks, which will be carried out simultaneously rather than sequentially.

First, we need to finish the digital scanning, image processing, and basic data entry for approximately 5,000 documents. These two groups of documents include many of the previously unreported documents from the National Archives and a miscellaneous group of documents for which scanning and data entry was postponed for a variety of reasons. Of these 5,000 documents that need to be processed, approximately 3,500 have already been digitized but have not had basic metadata added; the rest need both image processing and basic data entry.

Second (and most important and most time consuming) is the content data entry or what the project has called “Data Entry II.” 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. This 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, for it is this that 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 will be 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. Despite the substantial work in completing the scanning, and data entry for the remaining documents, the time involved in carefully assessing the work already complete and building and establishing a new office and staff, and the time-consuming nature of the content data entry, we estimate that we will still be able to complete Data Entry II for 3,200 documents during the period of this grant. Naturally, work will accelerate in the subsequent phases of the project, and we estimate the project will be fully completed in an additional four years beyond this grant period, if we receive adequate funding. Such estimates are, of course, based on what we have been told by ESU about the state of the collection and the time that they have estimated for different work processes. We will confirm this estimate against a full and careful audit of the collection that we will undertake as we shelve the project in its new GMU home.

An additional, important area of editorial work involves locating relevant documents that have not yet been added to the collection. The most time-consuming work involves a systematic search of era newspapers, which often published War Department documents, especially letters. Additional documents (some requested previously but never delivered; others we have more recently learned about) need to be gathered and scanned from the Massachusetts State Archives, the Historical Society of Pennsylvania, Harrington Library in Pittsburgh, and the Clements Library at the University of Michigan. In addition, the Continental Congress Papers need to be systematically searched for relevant War Department records. Because of the reduced funding for the project, we will postpone this work to a later phase of the project. Finally, because this project will extend beyond the proposed grant period and because we are planning on making this material permanently available, it is important to say something about the

long-term sustainability of the project. GMU is making a substantial cost-sharing commitment to the project both during and after the nineteen-month proposed funding period. That includes not only the formal cost sharing shown on the budget (e.g., staff time and indirect costs) but also free office space, free server space and website maintenance, and free storage space for documents in the GMU new Research building. The Center for History and New Media is also committed to the project, and it has a \$2 million endowment that provides a long-term stable funding base for its operations. Even when the project is fully complete and there are no more grant funds, we are committed to keeping the website running. As part of our long-term planning for the project, we will also begin addressing issues of digital preservation. Fortunately, CHNM's participation in the Library of Congress's NDIIP project puts in close touch with the latest developments in digital preservation. But in addition to working to insure the permanence of the digital files, we will consider creating a digital microfilm of all of the project files as a back-up copy.