6-30-2015

Interim Report, HD-51897-14, Image Analysis for Archival Discovery (Aida), June 2015

Elizabeth M. Lorang
University of Nebraska-Lincoln

Leen-Kiat Soh
University of Nebraska - Lincoln

Interim Report
HD-51897-14
Image Analysis for Archival Discovery (Aida)
Elizabeth Lorang
Leen-Kiat Soh
University of Nebraska-Lincoln
2015-06-30

In the second six months of work on "Image Analysis for Archival Discovery," the project team has continued making strides toward our goal of analyzing more than 7 million newspaper pages in Chronicling America for poetic content. We have hit a few challenging areas in our research and development work, and our work plan has shifted in some ways from that originally set out in our application, but we have implemented these changes with the fundamental goal of performing the major research outlined in our proposal—exploring image analysis as a methodology for discovery in digitized collections of historic materials via a case study of identifying poetic content in historic newspapers.

Activities undertaken from December 2014–May 2015:

- Development of an article describing the creation of our classifier for recognizing poetic content in historic newspapers; accepted and forthcoming in July/August 2015 *D-Lib* (completed)
- Development of Python program for parsing Chronicling America JSON files and batch retrieving JPEG 2000 image files (completed)
- Operationalization of entire process, from image retrieval to image processing, including moving to server environment (in progress)
- Development of project documentation (completed to current stage)
- Processing and classifying all Chronicling America images from the period 1836-1840 as a test case (in progress)
- Communication of results with relevant audiences, such as at the American Literature Association conference and via project website (completed)
- Pursuit of external partnership and additional source of funding: submission of Google Faculty Research Award application (completed; decision pending)

We are now in the process of retrieving large batches of images from Chronicling America and processing them with our software, which first segments the full page images into image snippets, reduces noise in the image snippets, and classifies the snippets as containing or not containing poetic content. We aim to complete work on approximately 25,000 page images from 1836-1840 in July.

Unfortunately, this deadline is far behind where we would like to be in order to process a full 7 million pages by the end of the grant period, and we have realized we were over-ambitious in imagining the ability to process the 7 million pages during the grant period. For example, in a recent phone call with Library of Congress staff to troubleshoot some issues we were having with retrieving images, we learned that if we were to retrieve all 10 million page images currently available via Chronicling America, this would be a 16-week process, if we were retrieving images constantly with no down time. This phone call with the Library of Congress also confirmed that our project is the first they have encountered that seeks to make use of large

numbers of Chronicling America's images. As a result, we have not benefitted from the readily available, pre-packaged bulk downloads for Chronicling America's textual data. For the start-up stage, however, we believe we can adequately test the principle methodology on a smaller subset of images and still understand necessary directions for future work. In addition, once we have the system fully operationalized—of which we are confident—and images retrieved, we can continue the processing beyond the end of the grant period.

Because we have been delayed in getting to large-scale deployment, we have not been able to be as aggressive as initially imagined in seeking external partnerships. We have, however, reached out to Google in a funding competition, and we have been in touch with the Library of Congress about our work. At this stage, forming an external partnership would be premature, but forging such relationships remains a central goal of the project team. We have instead focused outreach and dissemination efforts on sharing project work, both in our forthcoming *D-Lib* article and in conference presentations. In the first six months of the grant, team members presented at Digital Humanities 2014 and the Digital Library Federation Forum, and in the second six months of the grant, PI Lorang presented on the team's work at the American Literature Association conference. Therefore, while we are behind in and have had to reshape our vision of some project activities, we are ahead in others.

In addition, the project has also developed a significant dimension as a training opportunity for undergraduate students. Because we are working with undergraduate researchers on this project, we have had to spend significantly more time on instruction for all parts of the project, from developing code, troubleshooting problems, writing documentation, and writing up results. We believe that the opportunity for the students to work on an interdisciplinary research project at this stage in their education, and the benefits to us of teaching them all of these different aspects of project development, is a significant component of the project that we did not imagine or document in our original NEH proposal.