1-2016

Interim Report, HD-51897-14, Image Analysis for Archival Discovery (Aida), January 2016

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
2016-01-22

In the third six months of work on "Image Analysis for Archival Discovery," the project team has made progress toward the goals outlined in our report from June 2015. As we reported in June 2015, we realized that our original plan to analyze 7 million pages from Chronicling America was overly ambitious for the grant period, and we revised our goal to complete a thorough case study of our methodology and code for all newspaper images in Chronicling America from the period 1836-1840. Activities undertaken, toward this and other grant goals, from June 2015–December 2015:

- Publication of the article, "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections," *D-Lib Magazine* (July/August 2015). doi: 10.1045/july2015-lorang
- Completed pre-alpha version of complete code base
- Continued development of project documentation
- Made progress in case study of images from 1836-1840
- Developed partnership with a researcher from the University of Virginia
- Discussed future project directions with Institute of Museum and Library Services staff
- Filed for no-cost extension to continue grant work through June 2016

Following publication of our article in *D-Lib*, we heard from multiple parties about our work, including expressions of support for investigating this methodology, expressions of interest in collaborating, and expressions of interest for funding. We are pursuing research collaborations and are preparing to seek funding for additional stages of work from various funders.

We have also completed a first pass of analyzing images from 1836-1840. Based on our initial work with the 1836-1840 images, it's now technically possible for us to process 7 million page images (pending the time to retrieve all of the images from Chronicling America), but if we dove into processing 7 million pages, this approach would trade quality for quantity and would be at the expense of a more thorough understanding of the opportunities and limitations of the approach. In addition, we believe focusing on the case study model will better position us to extend the methodology and our code to other projects. In a first pass of analyzing images from 1836-1840, for example, we have identified a number of issues that create challenges for our current system, even among a set of newspapers relatively similar in layout and design. We want our system to be better able to accommodate a more diverse set of images as well as to handle noise introduced at various points in the printing, microphotography, and digitization processes. Further, as we indicated in our previous report, 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. It is more crucial at this stage, then, to focus on active research and development work during the grant period.

As we reported in June 2015, a significant part of Aida has been its role as a training opportunity for undergraduate students. Undergraduate students were co-authors of our *D-Lib* article and contributed to the project in significant ways, and they have leveraged their research experience from the project in coursework and internship opportunities. We have reached a point in the project development, however, where the co-PIs need to shift the entirety of their grant efforts to active research and development rather than training, in part because the needed image processing work was too advanced for the undergraduate researchers with whom we had been working. The research question has proven to be more complicated than we originally believed, and we require more time to complete the case study that is central to evaluating the efficacy of the proposed methodology. In particular, we have had to spend more time than originally accounted for in the project timeline refining our algorithms to deal with the inconsistencies in the input images. As a result, we filed for and received a no-cost extension to continue the grant period through June 2016. Going forward, we hope to find ways for both undergraduate students and graduate students to participate in the research process.