# White Paper

Report ID: 111870

Application Number: HD-51897-14

Project Director: Elizabeth Lorang

Institution: University of Nebraska, Board of Regents

Reporting Period: 6/1/2014-6/30/2016

Report Due: 9/30/2016

Date Submitted: 10/7/2016

White Paper
HD-51897-14
Image Analysis for Archival Discovery (Aida)
Elizabeth Lorang
Leen-Kiat Soh
University of Nebraska-Lincoln
2016-10-03

NB: This document was also submitted as the Final Performance Report for grant HD-51897-14, in accordance with NEH Performance Reporting Requirements as revised January 2013. The only difference between the Final Performance Report and this document is this cover sheet and the first footnote below.

## Project Activities

With its Office of Digital Humanities Start-up Grant, the Image Analysis for Archival Discovery (Aida) team set out to further develop image analysis as a methodology for the identification and retrieval of items of relevance within digitized collections of historic materials.[1] Specifically, we sought to identify poetic content within historic newspapers, using Chronicling America's newspapers (http://chroniclingamerica.loc.gov/) as our test case. The project activities we undertook—both those completed and those in process—support this goal and align well with the activities proposed in our original funding application and as approved by NEH. To achieve our goal of creating an image processing-based system to identify poetic content in historic newspaper collections, however, we also made strategic decisions along the way that shifted some of our efforts from those we initially planned when we drafted our funding proposal three years ago.

During the grant period, the Aida team developed, trained, and tested a machine learning classifier that can identify poetic content in pages of digitized historic newspapers based only on visual signals. We published early results of this work in *D-Lib Magazine* in summer 2015. We have since undertaken a detailed case study that tests the application of our classifier and methodology to a test set of more than 22,000 newspaper page images from the period 1836-1840. Significantly, we shifted our emphasis from processing *all* pages from Chronicling America to conducting this thorough, critical analysis and case study. This shift in plans corresponds with our desire to explore image analysis as a methodology for connecting users of digital archives with materials of relevance.

Upon developing the classifier described in our *D-Lib* article, we *could* have moved on to deploying the classifier across 7 million or more pages from Chronicling America as we initially planned. Taking this approach, we *could* then have focused on all of the poetic content that the classifier helped us to identify and started to do some analysis of the poetic content itself. Taking this approach, we would have both identified a significant number of poems at scale and would have been able to make some claims about the poems we found. Rather than focusing only on the subset of poems we would have been able to identify, though, we opted instead to undertake a careful analysis of the method itself in order to fully understand the potential for this methodology and its application to other scenarios. One reason for this change in approach was the tremendous variety in newspapers, both in their original formats and in variations caused by subsequent duplicative processes—both microphotography and digitization. This shift in emphasis required that we study the classifier and its results in detail—not simply use the classifier to find as many poems as possible with the classifier as it currently stands. Simply put, we resisted the urge to move on to quickly applying the classifier without fully understanding both the strengths and weaknesses of the classifier and our overall approach. We opted to undertake a case study that deals with a significant volume of material (22,000 pages), but a volume of material that we can also manually corroborate in order to develop a critical understanding of the system. See the Accomplishments section for further details.

With regard to publicity and promotion, the research team and the University of Nebraska-Lincoln (UNL) publicized both the grant award and outcomes of our work at strategic opportunities. University Communications at UNL prepared a press release shortly after the award was made. This press release was picked up in various venues, including by our local NPR affiliate who aired a story on our project, and by the Poetry Foundation's blog, *Harriet*. In addition, the UNL Office of Research and Economic Development featured our work in its 2014–2015 annual report and developed a promotional video about our project. We also presented on our work at multiple professional conferences to help publicize our work and its results. Conferences included Digital Humanities 2014 in Lausanne,

[1] This document was also submitted on October 6, 2016 as the Final Performance Report for grant HD-51897-14, in accordance with NEH Performance Reporting Requirements as revised January 2013.

Switzerland; the 2014 Digital Libraries Federation Forum; and the American Literature Association annual conference in 2015. Research team members also shared our work at many local events for a range of audiences. We developed a project website, http://aida.unl.edu, which serves as a portal to our research products and to publicity about our work. Throughout the grant period, we made our interim performance reports public, both depositing them in our institutional repository and linking to them from our project website. Finally, our article in *D-Lib*, "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections" (July/August 2015), marked a key effort in publicizing outcomes of our work. This article raised further interest in our work and led to a new collaboration with the University of Virginia for the next stage of our project.

## Accomplishments

When we prepared our grant application in 2013, we stated our final product and dissemination strategy: "[Our] work plan will result in a program for retrieving image content from Chronicling America via the site API; source code for processing newspaper images to identify content based on feature recognition; and a white paper for NEH assessing this methodology for historic newspaper research and research in digitized collections more broadly. We will also create a catalog of all pages from Chronicling America that include poetic content. To the extent that we are able, we will provide even more granular information, such as tying particular zones of an image identified as poetic content to the OCRed text. Project team members will pursue appropriate conference and publication venues for disseminating this work, as well as work with university communications to prepare a press release at the time a grant is awarded and at the project's completion." As described above, however, we shifted our effort from processing the entirety of Chronicling America (or the 7 million pages it was approaching when we prepared our application) to a focused case study of the period 1836-1840. We shared this shift in strategy in our interim report from January 2016. The goals for the case study were: first, scale the system; second, test and analyze the accuracy of the classifier on a large set of data. We focused on the five-year period 1836-1840 because it then represented the first five years' worth of content in Chronicling America, and we knew that we could expect to find a significant amount of poetic content in newspapers from this period.

The two tables below provide further details on our accomplishments and activities from the grant period. Table 1 summarizes at a high level the project activities undertaken during the grant period in relation to the deliverables outlined in our original funding application. Table 2 outlines actual accomplishments and offers a description of each accomplishment and the work undertaken/completed.

| Deliverable from Grant Application | Status | Comments |
|---|---|---|
| Program for retrieving image content from Chronicling America | Complete, final documentation in process | To be made available via GitHub repository by December 31, 2016. |
| Source code for processing newspaper images to identify content based on feature recognition | Complete, final documentation in process | To be made available via GitHub repository by December 31, 2016. |
| White paper for NEH | Complete | Submitted to NEH October 2016; deposited to UNL institutional repository and made available via project website October 2016. |

| | | |
|---|---|---|
| Catalog of all pages from Chronicling America that include poetic content; to the extent possible provide granular information, such as tying particular zones of an image identified as poetic content to the OCRed text | Will not complete at this time | Shifted emphasis to highly detailed case study of period 1836–1840. Change of plans discussed in interim report from January 2016. |
| Conference presentations and publications | Complete and ongoing | Presented at multiple conferences, with one major publication on the project from the grant period. See Table 2. |
| All source code made freely available via GitHub | In process | Currently in a private GitHub repository. Repository will be made public by December 31, 2016. |
| Reports and publications deposited in institutional repository | Complete | See Table 2. |
| Project image data and metadata deposited in UNL's data repository | Incomplete | To be deposited by December 31, 2016. |

**Table 1**: Comparison of deliverables and dissemination strategy from original funding application to actual project activities.

| **Accomplishment** | **Description of Accomplishment/Work** |
|---|---|
| Developed program for retrieving image content from Chronicling America | • Created a python script for batch retrieving images from Chronicling America. Currently, the script allows the user to define a temporal range for the images retrieved. This script can be used alongside our feature extraction and classification code or independently to allow users to undertake other kinds of work with images from Chronicling America.<br>• Retrieved 21,722 page images from Chronicling America for the period 1836–1840 |
| Developed code for segmenting page images from Chronicling America into page image snippets for use by feature extraction code | • Created program to evaluate a page image from Chronicling America according to several rules to determine whether it can be adequately segmented into snippets. If the image passes the rule checks, the page image is segmented into overlapping snippets for use by the feature extraction code.<br>• Segmented automatically all pages that passed segmentation rules (8,845 of 21,722 page images from Chronicling America passed the segmentation rules)<br>• Created 509,338 image snippets<br>• Completed analysis of the page images that passed and failed the segmentation rule checks. Determined qualities that affect our ability to segment pages. Analysis complete and to be published with next article on our work. |
| Developed a classifier for evaluating images from historic newspapers to determine | • Created training sets of images<br>• Established characteristics to use for feature extraction and classification |

| | |
|---|---|
| whether or not the images contain poetic content | • Described those characteristics as measurable features; developed algorithms for measuring features<br>• Trained classifier with labeled images and feature values, to help classifier learn what features tend to signal poetic content<br>• Tested classifier on training images<br>• Published results in *D-Lib Magazine*<br>• Tested classifier on Chronicling America page images from 1836–1840 that passed segmentation checks and segmentation (8,845 page images and 509,338 page image snippets) |
| Evaluated machine learning classification of 1836–1840 images (in process) | • Manually evaluated 8,845 page images as containing or not containing poetic content<br>• Used page-level manual classification to classify all snippets from false pages as false<br>• Manually analyzed the results of image segmentation on 144 newspaper pages and inventoried typical problems encountered by the current image segmentation technique<br>• In process of evaluating 93,627 remaining snippets as containing or not containing poetic content<br>• *To be completed*: analysis of results from comparing manual classification to machine classification and all features that affect classification accuracy. Compare results among newspapers published in languages other than English and English-language papers (November 2016) |
| Disseminated research results and publicized project | • Presented at professional conferences including Digital Humanities 2014, Digital Library Federation Forum 2014, and American Literature Association 2015<br>• Presented to local audiences, including faculty and students. Presentations included those at the 2014 Association for Computing Machinery Regional Programming Competition and a 2014 Faculty Fellows Forum at the University of Nebraska-Lincoln<br>• Created project website, http://aida.unl.edu<br>• Established GitHub repository for sharing code and documentation (currently private, will be public by end of December 2016)<br>• Deposited interim performance reports in UNL institutional repository, http://digitalcommons.unl.edu<br>• Participated in interview about project with local NPR affiliate, NET Nebraska. A story about our project aired in Fall 2014 (http://netnebraska.org/article/culture/943643/how-find-poem-200-year-old-newspapers)<br>• Published results of the first stage of our research in *D-Lib Magazine* (July/August 2015)<br>• Invited to featured project in UNL Office of Research and Economic Development annual report for 2014–2015 |
| Developed new collaboration | • Developed collaborative partnership with John O'Brien at the |

| | |
|---|---|
| | University of Virginia and the Institute for Advanced Technology in the Humanities at the University of Virginia<br>• Included as research team and sub-award applicant on a proposed/pending international grant competition |
| Sought additional funding for future stages of project work | • Applied for Google Research Award funding (Spring 2015) (not awarded)<br>• Applied for NEH Humanities Collections and Reference Resources (Summer 2015) (not awarded)<br>• Applied for IMLS National Leadership Grant research project funding (awarded, project formally begins December 2016) |

**Table 2**: Outline of major accomplishments and work completed as part of each accomplishment.

Between now and the end of November 2016, we will complete the case study using the research time of co-PIs Lorang and Soh. By the end of December 2016, we will have made available project data and source code from the NEH-funded work. We will then take up the next stage of our project, funded by IMLS, titled "Extending Intelligent Computational Image Analysis for Archival Discovery."

## Audiences

The primary audience for our project is information professionals in academic and research libraries. Other audiences include computer and information scientists and humanities researchers. We reached these audiences in a variety of ways, including through our public website and publicity efforts as well as through presentations at professional conferences and through our article, "Developing an Image-Based Classifier for Identifying Poetic Content in Historic Newspapers."

Our interim performance reports, available via the UNL Libraries' institutional repository, have been downloaded dozens of times, and our *D-Lib* article has already been cited at least one time (Padilla, 2016). Beyond these quantitative values, however, we have further evidence that our article and our work reached its target audiences. Following the publication of the article, a scholar at the University of Virginia reached out to us about a potential collaboration, and we also heard from a program officer at IMLS encouraging us to consider applying for IMLS funding. Both of these developments, which emerged from the *D-Lib* article, were crucial to our application to IMLS for further funding of our research endeavor.

## Evaluation

We have several ways to evaluate our work at this stage. We might evaluate the project as a whole, according to the goals, deliverables, and work plan that we created three years ago at the time we applied for funding: did we achieve the goals we set out for our project at that time? We are confident that we have achieved the highest-level goal of advancing image analysis as a methodology for identification and discovery in digital collections. In prompting and promoting conversation around this topic, our project has been successful, and we have also advanced the technical framework for doing this work. At times, the project certainly felt overly ambitious for the amount of time and support we had, and we did have to scale back some of our original ideas, such as thinking we'd be poised to start tying image classification data to the underlying structured textual data or that we'd be in a position to start pulling in and processing content from additional collections beyond Chronicling America. Ultimately, though, where our end project results differ from those we imagined, the differences have

emerged because of a critical re-evaluation of the work we determined to be most crucial to advancing the main project goal. This re-evaluation led us to work more intensely and critically but at a slightly smaller scale.

Another way to evaluate our project is by the quantitative measures of accuracy that we reported in our *D-Lib* piece on developing our classifier. At the classifier training stage, we achieved precision and recall of 90.58% and 79.40%, respectively (precision measures how many times we are correct whenever we predicted that an image contained poetic content; recall measures how many times we correctly identified poem snippets out of all poem images). These numbers dropped when when we tested the classifier—to 74.92% and 61.84%—but they remained encouraging enough that we decided to undertake a larger case study, as we did not want continued refinement to reflect only the small sample of our training and testing data. Once we complete the manual classification of the remaining 93,000 image snippets, our next step is to compare the machine and manual classifications, evaluate statistics for the 1836–1840 images, and then to look for appropriate places to revise our code and/or our methodology to achieve even better results.

Finally, we might evaluate our project according to "softer" outcomes. Two undergraduate students played a key role on the project during the grant period. They gained experience in research practices, algorithm development, and programming languages, and they contributed as co-authors on both internal and external documents. These students have moved on to internship programs, and our project played an important role in their early career development. Their contributions to the project, as well as those of other undergraduate students at an earlier stage of the project, have inspired us to pursue the creation of interdisciplinary laboratory at UNL for undergraduate students to continue to engage this project from a variety of angles and disciplines. Our planning for such a laboratory is in the germinal stages. The project also led to a new collaboration between the UNL research team and one at the University of Virginia. Similarly, we were asked to join an international research team on a pending project proposal, where our image-based methods will complement text-based analyses of digitized newspaper corpora. These emerging collaborations provide some evidence for the success of our project to date as well as its potential.

## Continuation of Project

We will wrap up the case study portion of our work by the end of November 2016. This remaining work is supported by the research time of co-PIs Lorang and Soh. In addition to concluding the case study, we plan to draft and submit a second article for publication out of our NEH-funded work and plan to complete the draft by the end of December 2016. This article will detail efforts to automatically segment newspaper pages and the challenges therein as well as a detailed analysis of our classification results. With the completion of this work, the team we will move on to the next stage of project development, which will be supported by IMLS. This next stage of our work aims to provide 1) open source software designed to process digital images of newspaper collections in multiple languages, including algorithms for visual feature extraction of poetic and advertising content; 2) datasets representing these visual features; 3) datasets documenting where the content appears in the newspapers; 4) an open access collection of poems extracted from newspapers; 5) a series of open access reports on our processes and methodologies; and 6) several training opportunities for members of library and archive fields to learn about our methods and how to use the software.

## Long Term Impact

We are beginning to get a sense of the potential long term impact of our work. Recently, co-PI Lorang presented on the Aida team's work at the inaugural Collections as Data symposium at the Library of Congress. We received considerable positive feedback about the way our work reframes approaches to digital collections in considering the value of digital images of textual materials as well as about the critical and evaluative approach we are taking to studying our classifier, its results, and the overall methodology. Furthermore, we described the results of our work to date and future plans as part of our application to IMLS for National Leadership Grant funding. Through two stages of peer review, reviewers emphasized the potential impact our research stands to have on digital collections and their users. Our approach stands out in terms of its implications for dealing with scalability and large datasets, and its flexibility in leveraging visual cues—instead of textual information—towards identification of certain types of texts in digitized collections.

## Grant Products

Grant products produced during the course of the project include the project website, data sets, interim reports, and a publication. Grant products are located in one of several repositories, depending on the nature of the product. See the table below (Table 3) for the location of the products. In addition, materials currently available are linked to from the project website (http://aida.unl.edu), and we will continue to update the project website to include links to new content.

| **Grant product** | **Where available** |
|---|---|
| Project website | http://aida.unl.edu Also serves as a portal to access all grant products below |
| Interim and final reports | UNL institutional repository, http://digitalcommons.unl.edu |
| Data sets | UNL data repository, https://dataregistry.unl.edu/ (to be deposited by end of December 2016) |
| Source code | GitHub, https://github.com/CDRH/aida (currently private; public in December 2016) |
| Publications | UNL Libraries' institutional repository, http://digitalcommons.unl.edu |

**Table 3**: Grant products and their location