3-2019

Interim Performance Report, LG-71-16-0152-16, Extending Intelligent Computational Image Analysis for Archival Discovery, March 2019

Elizabeth Lorang
University of Nebraska - Lincoln

Leen-Kiat Soh
University of Nebraska - Lincoln

John O'Brien
University of Virginia

INSTITUTE of Museum and Library SERVICES

INTERIM PERFORMANCE REPORT

For Projects with Award Dates after October 1, 2015

Please consult the IMLS Interim Performance Report Line Item Instructions when filling out this form.

1. Federal agency and organization element to which report is submitted: **Institute of Museum and Library Services**	2. Federal award or other identifying number assigned by federal agency: LG-71-16-0152-16	Page 1	of 6 Pages
		3a. D-U-N-S® number: 555456995	
		3b. EIN/TIN: 470049123	
4. Recipient organization (name and complete address, including ZIP+4/postal code): Board of Regents of the University of Nebraska 151 Prem S. Paul Research Center, 2200 Vine Street Lincoln, NE 68583-0861		5. Recipient identifying or account number: 25-1620-0028-001	
6a. Award period of performance start date (MM/DD/YYYY): 12/1/16	6b. Award period of performance end date (MM/DD/YYYY): 11/30/19	7. Reporting period end date (MM/DD/YYYY): 11/30/18	
8. Project URLs, if any: http://projectaida.org https://github.com/ProjectAida		9. Report frequency: ☒ annual ☐ semi- annual ☐ quarterly ☐ other If other, describe:	
10. Other attachments? ☐ Yes ☒ No Contact the appropriate IMLS program office to receive instructions for transmitting additional attachments.			
11a. Name and title of Project Director: Elizabeth Lorang Associate Dean		11b. Telephone (area code, number, extension): 402.472.2516	
		11c. Email address: liz.lorang@unl.edu	
12. Certification: By submitting this report I certify to the best of my knowledge and belief that this information is correct and complete for performance of activities for the purposes set forth in the award documents.			
13a. Signature of Authorized Certifying Official:		13b. Date report submitted (MM/DD/YYYY):	
13c. Name and title of Authorized Certifying Official: Jeanne Wicks, Director Office of Sponsored Programs		13d. Telephone (area code, number, extension): 402-472-3171	
		13e. Email address: Jwicks2@unl.edu	

Burden Estimate and Request for Public Comments: Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the Institute of Museum and Library Services, 955 L'Enfant Plaza North, SW, Washington, DC 20024-2135.

The purpose of the Interim Performance Report is to provide a record of grant-funded project activities at annual intervals throughout the grant period. If you have questions concerning the interim performance reporting requirements, you may address them to the Program Officer assigned to your grant and whose name and contact information appears in your Official Award Notification. IMLS may share Interim Performance Reports with grantees, potential grantees, and the general public to further the mission of the agency and the development of museum and library services. Reports may be distributed in a number of ways and formats, including online.

14. Recipient Organization:

Board of Regents of the University of Nebraska

15. Project Title:

Extending Intelligent Computational Image Analysis for Archival Discovery

16. Project Summary:

The primary goal of "Extending Intelligent Computational Image Analysis for Archival Discovery" is to investigate the use of image analysis as a methodology for content identification, description, and information retrieval in digital libraries and other digitized collections. Building on work started under a National Endowment for the Humanities' Office of Digital Humanities Start-up Grant, our IMLS project seeks to 1) analyze and verify our previously developed image analysis approach and extend it so that it is newspaper agnostic, type agnostic, and language agnostic; 2) scale and revise the intelligent image analysis approach and determine the ideal balance between precision and recall for this work; 3) distribute metadata and develop a new digital collection using the extracted content; and 4) disseminate results, including adding to the scholarly literature on these topics and providing training for members of library and archive communities.

In the second year of the project, the Aida team made considerable headway in the goals of our grant. While we have continued to focus exclusively on poetic content to this point, year two was an important year for assessing the efficacy of the approach and extending it such that it might be newspaper- and language-agnostic. In addition, we assembled a large set of data and evidence to help us consider the balance of precision and recall as well as to consider revisions to the overall approach given what we're learning in this area. We also have a functional metadata model and have made major steps toward developing a new digital collection out of the poetic content observed during the project, and for distributing metadata about the content. Finally, team members shared about the work at four major conferences, to audiences of digital library professionals and specialists and literary scholars. Team members prepared three publications, which are currently out for review, a detailed report analyzing the extension of the approach to a new corpus and have generated notes toward additional articles and other writing for year 3.

17. Activities

Activities Proposed in Your Application	Activities Completed during the Reporting Period	Explanation of Any Variance
Prepare first open access report documenting success and challenges of year one work (All)	Published on project website the full-text and slides of two major presentations and the	We had begun substantive work on a report focused on machine learning and digital library development, however, we

	text of a third more minor presentation, all related to key aspects of the project.	realized that the content of these various presentations were substantive interventions into the conversations on machine learning and digital libraries in their own right and drew on some of the research we had been doing. In order to expedite getting this information circulate, we prepared the text of these publications for distribution. They are hosted in the University of Nebraska-Lincoln University Libraries' institutional repository and made available there, and also from the project website.
Develop preprocessing approaches to accommodate a greater variety of newspapers (Soh; CSE GRAs)	Tested and refined pre-processing approaches on the Burney Collection of British Newspapers and on newspaper pages from several other corpora, including from the Internet Archive.	
Present on work and/or lead a workshop at Code4Lib 2018 (Lorang; Soh; and/or DH GRA, UNL)	We did not present on our work at Code4Lib, however project team members presented an invited keynote at the 2018 HathiTrust Research Center conference; participated in a panel at the Joint Conference on Digital Libraries in June; and delivered an invited opening presentation at the National Digital Newspaper Program meeting in September.	Opportunities other than Code4Lib emerged for presentation, including to audiences that seemed ideal for the purposes of our work.
Present on work and/or lead a workshop at ASECS 2018 (O'Brien; DH student, UVA)	Presented on the project at ASECS 2018, as planned.	
Prepare "ground truth" datasets for advertisements from Chronicling America and the Burney Collection; document relevant features of interest (Lorang; DH GRA, UNL; O'Brien; DH student, UVA)	Prepared larger ground truth sets focused on poetic content.	Because our work on poetic content is taking longer than anticipated, and we continue to test and refine approaches for poetic content, we have postponed treating other type of generic content at this point. In addition, we have postponed this work because we are developing some new approaches to segmentation that would have changed the type of ground truth set we needed. Rather than do this work multiple times, we have postponed it for now, but anticipate taking it up in year 3.

Continue design, development phase of database of poems (O'Brien; DH student, UVA; IATH team)	Completed initial design of database, as well as of metadata model. Began inputting poems into database, to test the technical and metadata infrastructures.	
Convene monthly conference calls with advisory board (All)	Convened 2 meetings of the advisory board over the 12-month period.	In this active development stage, we were not finding that we had specific questions and concerns to take to the advisory board. We wanted to make the most of their time, so we did not convene the board if we did not have specific agenda items to discuss.
Develop and test classifier for advertising content (Soh; CSE GRAs)	Not pursued.	As above, for developing the ground truth set, we have postponed this work due to spending additional time on poetic content and as we continue to refine and revise our overall approach. We want to bring the best possible approach to these other types of content, when we have a good overall system in place.
Deploy preprocessing approaches and full processing pipeline for poetic content on previously processed Chronicling America pages; analyze and verify results (Lorang; DH GRA, UNL)	Completed; publications submitted for review.	
Deploy preprocessing approaches and full processing pipeline for poetic content on previously processed Burney Collection pages; analyze and verify results (O'Brien; DH grad student, UVA)	Completed; report forthcoming as open access report for year 2.	
Hold project meeting and development sprint (All)	Project meeting and development spring was held in Washington, DC, in September 2018.	
Continue refining poetic content classifier as necessary (Soh; CSE GRAs)	This was a major area of activity in 2018, including exploring and testing alternative approaches to classification.	
Lead a workshop at Digital Library Federation Forum to train participants on the software and get community feedback (UNL team)	As above, for Code4Lib, we did not present at the DLF Forum, however project team members presented an invited keynote at the 2018 HathiTrust Research Center conference; participated in a panel at the Joint Conference on Digital Libraries in June; and delivered	Opportunities other than DLF Forum emerged for presentation, including to audiences that seemed ideal for the purposes of our work. In addition, we are not in a position yet to train others to use the software, as we continue to develop it. If we had unveiled it at DLF Forum, the software would have been buggy and confusing to use, and that reality could

	an invited opening presentation at the National Digital Newspaper Program meeting in September.	have negatively impacted overall perception and reception of our work.
Investigate strategies for sharing metadata with originating collections and strategies for desiloing the data (Lorang; O'Brien; DH GRA, UNL; DH student, UVA)	Work is underway, including with other projects focused on poetry as well as with the vendor of the Burney Collection of newspapers (Gale).	
Perform computational analysis of historic newspaper characteristics (postponed from year 1)	Develop and tested approach for analyzing full-page digital newspaper images, beginning with features such as bleed-through, orientation skew, range effect, and overall noise and density of content. This analysis will be helpful for understanding the papers/digitized images more broadly as well as for developing our computational approaches. Early results of this work are presented in the talk we gave at the National Digital Newspaper Program meeting.	

18. Changes

Type of Change	Description	Date of Approval (if applicable)

19. Lessons Learned

1. The significant variance in historical newspaper document images in terms of contrast and layouts, and also a wide range of noise effects (such as bleed through, range effects, skew orientations, and blobs) stress our original approach and require revisions to our algorithm implementation. Our original approach was tested on a subset of images from the Chronicling America repository and now after extending our approach to a considerably larger set, we learned that we had over-estimated the generalizability of our approach, even though we are still only looking at historical newspaper document images and poetic content. More specifically, with regard to one of the collections we've tested, we knew going in that the Burney Collection of Historic Newspapers would present greater problems than the Chronicling America collection because of the age of the source material and the period when the collection was digitized (much earlier than most of Chronicling America). But the problems with Burney are probably greater than anticipated. The poor quality of the images, owing either to issues with the original pages or their digital surrogates, means that at the moment, many pages are essentially unreadable by the software.
2. To deploy our prototype, in order to make it user-friendly, our software package needs to have better documentation (e.g., user manuals), especially on trouble shooting, and also tips on how to install and run the prototype on different operating systems or platforms.
3. Our solution's approach, that integrates image processing techniques and machine learning techniques, has now evolved into investigations into two relatively distinct areas: image segmentation or zoning to divide up a document image into separate zones for easier processing, and deep learning-based classification to avoid having to carry out extensive feature extraction. These two areas have emerged from our effort to extend our original approach. The zoning work will help us automate our approach to identify coherent image snippets of each newspaper page; while the deep learning using convolutional neural networks will allow the solution to be independent of design choices of feature extraction.
4. Standards for metadata for poetry are surprisingly underdeveloped. One contribution this project can make is to help advance the discussion about these standards, in part because the metadata issue is more pressing for retrieval and discovery of these items than for poems printed in codex volumes.