2-27-2018

Interim Performance Report, LG-71-16-0152-16, Extending Intelligent Computational Image Analysis for Archival Discovery, February 2018

Elizabeth Lorang
University of Nebraska - Lincoln

Leen-Kiat Soh
University of Nebraska-Lincoln

John O'Brien
University of Virginia

INTERIM PERFORMANCE REPORT

Please consult attached 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 (mo/day/yr):
12/1/16

6b. Award period of performance end date (mo/day/yr):
11/30/19

7. Reporting period end date (mo/day/yr):
11/30/17

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 IMLS program office to receive instructions for transmitting additional attachments.

11a. Name and title of Project Director:
Elizabeth Lorang
Associate Professor & Humanities Librarian

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:
Jeanne Wicks

13b. Date report submitted (mo/day/yr):
02/27/2018

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

14. Agency use only

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 who is assigned to your grant and whose name and contact information appear 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.

15. Recipient Organization

Board of Regents of the University of Nebraska

16. Project Title

Extending Intelligent Computational Image Analysis for Archival Discovery

17. 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 first year of the project, the Aida team's work has focused on several key areas, contributing to the goals and strategies outlined above: assessing and improving our algorithms and functional code for greater accuracy in classifying poetic content in historic newspapers; studying characteristics of historic newspapers across time-period, geography, and language; planning toward distributing metadata and developing a new digital collection; and distributing open code, datasets, and other information related to project work. Functionally, all of our efforts in year one map to goals and anticipated outcomes identified in our grant proposal, even when there is some variation between the activities we anticipated in the application stage and those undertaken in year one.

18. Activities

Activities Proposed in Your Application	Activities Completed during the Reporting Period	Explanation of Any Variance
Perform computational analysis of historic newspaper characteristics	• Revised and developed new approaches and algorithms for the overall classification scheme to improve algorithm robustness. • Developed tool for human visual comparison of different approaches to binarization and consolidation methods, to analyze and assess results. Conducted multi-person	We have postponed the computational analysis of historic newspaper characteristics until year two. Rather than undertaking this new area of inquiry, the team focused on documenting and generalizing the existing code and approach, and this work led to some research and findings that warranted further revision to the algorithms—and possibly to the overall approach. After training up to 17,000 image snippets, up to

	assessment of the approaches using the visual comparison tool.	10,000 episodes on the ANN, we did not see significant increase in the training accuracy. Conceptually, we should be able to train until it is over 90% accurate (training wise) even at the expense of over-fitting. That we were not able to reach such a high training accuracy indicates that this Back-Propagation Neural Network (BPNN) approach might be a dead-end and, in a larger sense, that this symbolic approach (extracting visual cues explicitly, modeling them as numeric values, training a BPNN) might also be a dead-end. At the same time, when we explored the Deep Learning approach, we achieved high training accuracy (> 90%) when we simply just fed the entire binary image snippets into training a Convolutional Neural Network (CNN). The decision to focus first on this area of investigation is the main reason for the variation from the proposed tasks or activities.
Reorganize existing code base to enable flexible experimentation and to streamline batch running processes	• Existing codebase reorganized, for both code management and development purposes. • Virtual server set up with Center for Digital Research in the Humanities and configured with enough storage to allow for maintaining digital images for batch running processes. • Explored possibilities for parallel processing on high-performance computing cluster at University of Nebraska-Lincoln.	None
Publish code to GitHub repository	Code published to repository using proper staging and publishing procedures.	None
Develop GitHub pages version of project website linked to code repository on GitHub	New project website developed on GitHub pages platform. In addition, the team created a GitHub Organization's page, which links to the project website and to the teams several code repositories.	None
Prepare "ground truth" poetry datasets from Chronicling America and the Burney collection	Prepared	None

Design, develop database of poems from Burney Collection	- Identified goals for database - Conducted comparative analysis of similar poetry databases - Focused advisory board conversation on desirable metadata fields and approaches for meeting those goals - Began development of database schema	We realized that a prior step was needed before design was undertaken, which involved comparing comparable poetry databases to identify key features that needed to be included and to allow for possible future interoperability. The initial database schema has now been designed, and we are going to begin implementing development and wireframing.
Hold project meeting and development sprint	- Project meeting held virtually. - Project development sprint replaced with other code development sessions.	See **Changes** below for further description of the changes. Ultimately, this change in location and process did not lead to a difference in the deliverables for the grant.
Analyze a minimum of 30,000 pages from Chronicling America; analyze and verify results and compare with results from NEH-funded start-up phase (Lorang; DH GRA, UNL)	- Analyzed snippets derived from 12,000 pages - Performed detailed comparison and analysis of results to start-up phase/results - Created reports documenting comparison/analysis - Developed new algorithms to respond to challenges and develop a more robust approach	Rather than focusing on analyzing the higher number of pages, we focused on performing a much more detailed analysis on a smaller number of pages. This analysis is the basis for future code developments. If we can improve our overall methodology at this stage, we stand to be able to process many more pages more effectively, than if we had focused on the number of pages from the outset.
Analyze a minimum of 5,000 pages from the Burney Collection (O'Brien; DH student, UVA)	Analysis in progress	The project team has not yet completed this analysis because we ran into some challenges deploying our new system in a different environment/with new users and with a collection of objects whose basic design (column width, numbers, page layout, etc.) differ in fundamental ways from the Chronicling America dataset. We are addressing the challenges and anticipate concluding this work in the near future. Identifying and addressing such challenges is a key component of assessing the viability of our overall methodology/approach. Currently, we are also assessing the feasibility/desirability of adding the Nichols Collection of early newspapers at the Bodleian to our dataset.
Convene monthly conference calls with advisory board (All)	At beginning of project, held monthly board meetings. Meetings now held on a quarterly basis.	Advisory board meetings are now quarterly, rather than monthly. The project team sends monthly updates to the advisory board in months that they do not convene. This difference in the frequency of board meetings from the initial plan is due both to scheduling for such a large meeting, of up to a dozen attendees, as well as to the pace of project work. Should more frequent board meetings become necessary in the future, we will adjust the meeting schedule and also

		pursue the possibility of advisory board subgroups around particular areas.

19. Changes

Type of Change	Description	Date of Approval (if applicable)
In-person, all-team meeting not held; virtual half-day meeting held in place.	Team members were not able to travel for or host an in-person all-team meeting due to family health emergencies. We had already made travel arrangements for the meeting and a portion of the pre-paid travel expenses were non-refundable. In lieu of this meeting, team members met virtually for a half-day session.	

20. Lessons Learned

1. Historical newspapers' qualities, as well as those of the digital images documenting them, are even more varied than we previously considered, and a one-size-fits-all solution might not be feasible. Should such a one-size approach not be feasible, we are interested in determining what parts of the methodology are generalizable, what factors affect generalizability/customization, where customization might be necessary, and the implications of this information for the possibility of wide-adoption of such image analysis in the development of digital libraries.

2. Symbolic, knowledge-based approaches to automated image classification might not be as robust as deep-learning-based approaches.

3. Robust page segmentation to identify textual components on a newspaper page is non-trivial.

4. The quality of digitization—affected by such factors as the time period in which the digitization was undertaken and influenced also by the qualities of the microcopies that serve as the basis of digitization, among other factors—introduces additional variables for evaluation and analysis at multiple points.