Abstract

This proposal seeks research grant funding in the amount of $462,318 to allow the Image Analysis for Archival Discovery (Aida) research team to investigate the use of image analysis as a methodology for content identification, description, and information retrieval in digital libraries and other digitized collections. Our approach leverages machine learning to build an intelligent computational system that extracts visual cues from image training sets, trains a classifier to recognize these visual cues, and subsequently deploys the system to process new digital images and identify similar content. Our project, titled "Extending Intelligent Computational Image Analysis for Archival Discovery," will result in the following products: 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.

Our work addresses the challenges of identification, description, discovery, and analysis within digitized collections, aiming to refine the ability for librarians, archivists, and researchers to identify and use particular items that they seek. Our project objective is to develop a methodology and corresponding software that will enable librarians and archivists to utilize image analysis for identifying, isolating, and making more readily available specific kinds of content in large digital collections of historic newspapers. To this end, we will focus on developing image analysis procedures and software to identify poetic content and advertisements in historic newspapers over two centuries and continents, representing multiple languages and types of newspapers. Our project will improve the library community's capacity to identify digital content at and below the item level as well as researchers' access to this content. In developing open source software and methodologies that will increase access to information currently obscured in large digital collections, our project falls under the IMLS National Digital Platform priority.

"Extending Intelligent Computational Image Analysis for Archival Discovery" will expand and improve access to digital content and services. We will achieve this through developing concrete products, including software, data sets, and new digital collections, as well as by developing conceptual approaches and methods. In particular, we will address a problem posed by the creation of many enormous digital collections in which various kinds of content are not disaggregated: how can we connect users to specific kinds of text and items of interest within the vast seas of large digital libraries of books, journals, newspapers, and more? We will address this problem by developing machine learning classifiers that identify textual content in historic newspapers based on the content's visual features; making the related software and algorithms freely available and licensed for reuse and continued development; linking and distributing extracted metadata about the items; adding new research partners to our team from the University of Virginia; offering training and workshop opportunities for learning about the methodologies and how to use the software; and developing new digital collections as a proof of concept of the types of work that become possible when we are better able to identify materials within disparate and mixed digital collections at scale. When fully realized, image processing approaches and intelligent classifiers, such as those we are developing, are likely to become part of the standard toolkit of archivists. Our methods and software offer a mechanism for more fine-grained description than is currently feasible. Our project thus also engages the IMLS question "What will move library and archival services in the United States forward?"

Statement of Need

This proposal seeks research grant funding in the amount of $462,318 to allow the Image Analysis for Archival Discovery (Aida) research team to investigate the use of image analysis as a methodology for content identification, description, and information retrieval in digital libraries and other digitized collections. Our approach leverages machine learning to build an intelligent computational system that extracts visual cues from image training sets, trains a classifier to recognize these visual cues, and subsequently deploys the system to process new digital images and identify similar content. We began this work at the University of Nebraska–Lincoln (UNL) with an NEH Office of Digital Humanities Start-up Grant and have shared preliminary results in "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections" (*D-Lib Magazine*, July/August 2015). With IMLS research grant funding, we seek now to develop a new partnership, extend Aida's software across a more diverse range of digitized newspapers and textual forms, and assess the broader potential of image analysis as a methodology for information classification, identification, discovery, and retrieval in digital libraries. Specifically, we propose to: 1) analyze and verify our 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.

Our project, titled "Extending Intelligent Computational Image Analysis for Archival Discovery," responds to a pronounced and growing need in the digital library community: for all the wealth of digitized primary materials now available online, finding materials of relevance in these collections remains difficult for all but the most routine use cases. Indeed, as the amount of material grows, so too do the challenges of locating relevant items. This challenge is documented most recently in Green and Courtney's (2015) "Beyond the Scanned Image: A Needs Assessment of Scholarly Users of Digital Collections." In their surveys and interviews of scholars working with digital collections of primary materials, the authors found that "One of the most prominent challenges cited by respondents in their use of digital collections was the inability to search effectively through the collection materials" (p. 695). One respondent told them, "The ability to conduct corpus wide inquiries is still severely limited. Search tools are good for finding needles in haystacks, terrible for extracting data for subsequent manipulations" (p. 696). These findings corroborate DeRidder and Matheny's (2014) findings about the difficulty of locating items of interest in digital collections as well as Strange, McNamara, Wodak, and Wood's (2014) findings about the importance of being able to distinguish between genres of text in newspapers for successfully analyzing the corpora. Underwood (2014) succinctly puts the problem this way: "Although methods of literary analysis are more fun to discuss, the most challenging part of distant reading may still be locating texts [within digitized collections] in the first place."

In developing open source software and methodologies that will increase access to information currently obscured in large digital collections, our project falls under the IMLS National Digital Platform priority. Our project will improve the library community's capacity to identify digital content at and below the item level as well as researchers' access to this content. Our work addresses the challenges of identification, description, discovery, and analysis within digitized collections, aiming to refine the ability for librarians, archivists, and researchers to identify and use particular items that they seek. We will develop computational approaches for identifying content at scale, particularly in digital libraries where both the collections and individual items in the collections are highly heterogeneous (e.g., newspapers, magazines, journals, scrapbooks, certain types of books). Researchers know that materials of interest to them are in these collections, but finding those materials can be difficult, both because of the level of description that the items have when they are made available and also because of the ways in which platforms allow users to query the data to find what they're looking for.

We believe that the library community can and should better leverage the information potential of digital images to aid in the classification and description of materials and to connect researchers with the resources they need. Images created in the digitization of primary materials contain a wealth of machine-processable information, and this information is significantly under-utilized.When fully realized, image processing approaches and intelligent classifiers, such as those we are developing, are likely to become part of the standard toolkit of archivists. Our methods and software offer a mechanism for more fine-grained description than is currently feasible. Our project thus also engages the IMLS question "What will move library and archival services in the United States forward?"

Both academic and industry sectors are increasingly interested in image-based methods for identification and discovery. One significant area of development is in image recognition and image searching of primarily visual works, including artwork and contemporary mass photography, such as photos posted to Flickr and Instagram. There is an increasing array of computer vision software designed for such applications, including OpenCV, Pastec (based on OpenCV), and Google's Cloud Vision API and related DeepDream software. Industry applications of these and related software include Google Goggles, Amazon Firefly, and Flickr's Magic View. Perhaps the most well known, popular application of computer vision is Facebook's facial recognition algorithm, which allows the platform to automatically tag people in photographs. In more academic settings, researchers are attempting to identify individuals in historic photographs as well as to detect prints made from the same woodblocks. In addition, the Software Studies Initiative at the CUNY Graduate Center has developed ImagePlot for exploring large collections of digital images and organizing images according to their visual characteristics. The Robots Reading *Vogue* project at Yale, for example, has used ImagePlot to study the "colormetric space" of the magazine's covers. Here again, the emphasis is on contemporary photography and also images of artwork. As we wrote last year in *D-Lib*, however, there has been far less work in investigating image analysis for primarily textual materials. Instead, image processing of textual materials has largely been regarded as a means to an end—as necessary for segmenting textual materials for optical character recognition (OCR)—rather than as a method of analysis in its own right.

Shifts in thinking are, however, emerging. In launching the Internet Archive's Book Images Project, Leetaru and Miller (2014) point out the limits of our collective imaginations with regard to image search, which has heretofore been seen as "a tool exclusively for searching the web [for images], rather than other modalities like the printed word" (n.p.). There is also an emerging body of work in the arts and humanities interested in the visual signals or visual analytics of textual materials. Recently, the Visual Page research team has analyzed visual features of pre-selected, single-author books of Victorian poetry in order to explore visual meaning in digital images of print pages. (Co-PI of the Visual Page project, Natalie Houston, served on the advisory board of our NEH Start-up Grant.) The Visual Page team has adapted an OCR engine to utilize the image segmentation done in OCR processes and then measure visual features of the segmented texts. In a similar vein, the shapes viewer at BloodAxe: The Poetics of the Archive (bloodaxe.ncl.ac.uk/explore/#/shapes), allows users to explore its archive by drawing the shape of a poem. The search then returns poems of similar shapes. As with the Visual Page, the BloodAxe project begins with known, pre-selected poems from a relatively small corpus. Nonetheless, it does begin to suggest alternate approaches beyond browsing and text-based searching to connect users with material of relevance in digital collections. Our project takes this work further to study visual signals of text and use them for content-based image identification within both heterogeneous collections and within individual items of heterogeneous, mixed genres and forms. Our research indicates that we are the first project to use image analysis for the purposes of generic identification in historic newspapers.

Concomitantly, researchers have begun to ask exciting questions of newspaper collections and to use these collections to make significant literary and cultural arguments. The Viral Texts project (viraltexts.org), for example, is mining nineteenth-century U.S. newspapers to find texts that "went viral," spreading across the country and internationally as local newspapers reprinted them, and then to theorize the characteristics of widely-disseminated items. Researchers in the School of English at the University of Sheffield have recently launched the Sheffield: Print, Poetry, and Protest project, which identifies and reprints political poetry published in radical newspapers in Sheffield during the period of the French Revolution and the Napoleonic Wars. Multiple projects in the United States have created collections of so-called runaway slave advertisements published in newspapers, including in Texas and North Carolina papers. These advertisements hold significant potential for conducting genealogical and family research, for augmenting historical scholarship, and for creating new works, both artistic and scholarly, that foreground the systemic impacts and human tolls of American slavery. Two such examples are the Twitter bots Texas Runaway Ads (@TxRunawayAds), which tweets runaway ads multiple times each day, and Every Three Minutes (@Every3Minutes), the latter which features images of advertisements for the sale of enslaved persons, among other archival materials. Researchers are poised, then, to ask new and potentially paradigm-shifting questions of newspaper and journal archives.

To do this work, however, users must be able to find materials of interest, and to be able to find such materials at scale, not only at the level of anecdote. They stand in need of methods and tools such as those the Aida project provides. Indeed, Ryan Cordell, a primary investigator on Viral Texts and member of Aida's NEH Start-up Grant advisory board, has shared that while the Viral Texts team has been pursuing generic identification via language models, the challenges of OCR create significant barriers to using language-based models for generic or formal identification. Furthermore, language-based models are by their nature language specific, while an image processing approach can be language agnostic. These concurrent developments—increasing challenges to finding items of interest in digital collections, the remarkable questions researchers aspire to pursue in these collections, and an emergent shift in thinking about the potential of image analysis—indicate that the time is right for the development of new, image-based approaches to working with digital archives. Our research is at the nexus of these developments and will enable the identification—and thereby the discovery and, ultimately, analysis—of the growing millions of items in digital libraries.

Impact

"Extending Intelligent Computational Image Analysis for Archival Discovery" will expand and improve access to digital content and services. We will achieve this through developing concrete products, including software, data sets, and new digital collections, as well as by developing conceptual approaches and methods. In particular, we will address a problem posed by the creation of many enormous digital collections in which various kinds of content are not disaggregated: how can we connect users to specific kinds of text and items of interest within the vast seas of large digital libraries of books, journals, newspapers, and more? We will address this problem by developing machine learning classifiers that identify textual content in historic newspapers based on the content's visual features; making the related software and algorithms freely available and licensed for reuse and continued development; linking and distributing extracted metadata about the items; adding new research partners to our team from the University of Virginia; offering training and workshop opportunities for learning about the methodologies and how to use the software; and developing new digital collections as a proof of concept of the types of work that become possible when we are better able to identify materials within disparate and mixed digital collections at scale. Significantly, our classifiers will work across a broad range of newspapers from the United States and Britain, spanning two centuries and multiple languages.

Thus far, the Aida team has focused on poetic content in newspapers, and we will continue this emphasis for part of the proposed grant work as well as move beyond poetry. Poetry in newspapers is visually distinctive and was widespread in eighteenth- and nineteenth-century British and U.S. papers. Many thousands of poems—some estimates say millions—were published in newspapers in this era. Most of these poems have not been accounted for in standard histories either of journalism or of literature. Given the visual distinctiveness of poetry and the immediate potential for scholarly impact, we seek to refine our methods on poetic content in newspapers but now also to cross geographic, temporal, and linguistic boundaries. Importantly, we will also extend our approach beyond poetic content and focus on a second type of textual content: advertisements. The reasons for moving on to advertisements are two-fold: 1) advertisements are an important area for cultural study in their own right; 2) sometimes advertisements are noise, and being able to filter them out is valuable. In either scenario, our work with advertisements is designed to provide users of these collections with greater means of access to content of interest, in the one case guiding users to the advertisements and in the other filtering out the significant noise that the voluminous advertisements in newspapers can create. Over time, our expectation is that further expanding Aida's software and related methods will allow users—whether librarians, archivists, or others—to isolate and recover many forms of content within historic newspapers, such as obituaries, marriage notices, recipes, shipping news, commodity prices, puzzles, and stock tables.

In subsequent stages of this work, we will seek to extend our image analysis beyond newspapers to other heterogeneous forms and collections as well. For example, in digital collections in which individual items are letters, diaries, or scrapbooks, classification via image analysis can play a significant role. Imagine being able to identify the contents of letters or diaries below item level, such that users might discover letters or diaries that include newspaper clippings, songs and verse, and other visual forms such as hand-drawn maps. Such discovery becomes possible not because of manually-generated and textually-encoded metadata but because of the potential of image analysis and machine learning classifiers to identify visual signals of interest. Ultimately, our work will make possible new strategies and workflows for librarians and archivists to identify and classify materials and will allow researchers in library and archive fields and beyond to pose new questions of digitized collections. In the long view, "Extending Intelligent Computational Image Analysis" has the potential to affect all users of digital libraries.

More immediately, this collaborative research will result in the following products: 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. We will make the software available so that users may process additional newspaper collections and also so that developers may further extend our methods and processes. Similarly, the feature extraction datasets will allow others to verify our work as well as to build on it in new—and potentially entirely different—ways, while the poem datasets will set the stage for corpus analysis and new collection building. To that end, we will also use poems identified with Aida's software to produce an alpha version of an open access database of poetic content gleaned from the newspapers, one that will allow users access to thousands of poems that we have identified and that will serve as an example of one kind of research outcome from the software's deployment. We will write a minimum of one open access report each year. These reports will not only document project work and outcomes but also will include bibliographies of relevant current scholarship. Finally, we will convene at least two training workshops at appropriate venues, such as Code4Lib and the annual Chronicling America meeting.

The completed software will be publicly available via our GitHub repository and will be issued under a GNU General Public License (GPL), version 3. GPL is a free software license that allows users to run, share, and modify the software. Data and metadata about the poems and advertisements are in the public domain and will be formally noted as such to encourage reuse. We will work with the UNL data curation librarian to identify appropriate data repositories. The web resource for the poems database will be open access and the data there also will be reusable by any researcher. The open access reports and training materials will be made available under a Creative Commons Attribution (CC BY) license. See the Digital Stewardship Supplementary Information Form for more on data retention and management and licensing.

To allow for input, consensus building, and buy-in from others in the field, we have assembled an advisory board of scholars and scholar-practitioners in library and information science and library services. Advisory board members are Paul Conway, Associate Professor, University of Michigan School of Information; Jody DeRidder, Head of Digital Services, University of Alabama Libraries; Adam Farquhar, Head of Digital Scholarship, British Library; Emily Gore, Director of Content, Digital Public Library of America; Patricia Hswe, Co-head, Publishing and Curation Services, Penn State University Libraries; Bethany Nowviskie, Director, Digital Library Federation; Ayla Stein, Metadata Librarian, University Library, University of Illinois at Urbana-Champaign; and John Unsworth, University Librarian and Dean of Libraries, University of Virginia. This advisory board brings to the project expertise in digitization practices and standards; studies of discoverability and usability of digital collections; digital scholarship and digital collection services; metadata exchange, including linked data; and national/international digital library infrastructure. The advisory board will play an active role in our project, and we will consult with them on a monthly basis.

The board will be one of the chief mechanisms for community input, but team members also will engage communities of practice at conferences and workshops including at Code4Lib, the Digital Library Federation Forum, and the Modern Language Association annual meeting, among others. We also been in contact with Chronicling America and of Gale, the developers of the two main digital newspaper collections we will use for our work, and we will seek their feedback. Gale has provided a letter of support, and we have had several conversations with members of the Chronicling America team. While they cannot make commitments beyond what they would offer any researcher, they will help us acquire data and discuss theoretical uses within their capabilities. In addition, we will be invited to present a workshop on the software as part of the 2017 or 2018 Chronicling America meeting program. Furthermore, team members' backgrounds in library and information science, literary studies, and computer science will enrich the perspectives brought to the research questions as well as provide opportunities for multidisciplinary engagement. Throughout, we will seek many opportunities for feedback from advisory board members and and our larger professional communities, those of libraries and archives professionals and researchers, humanities researchers, and computer science researchers.

We plan on deploying a variety of quantitative and qualitative measures to evaluate our work. First, and most obviously, did we address each of the four objectives set out in this proposal (see Project Design below)? A second, more finely calibrated metric will be a quantitative evaluation of just how successful our method is for identifying specific generic content in historic newspapers, including evaluating precision and recall and calculating corresponding F-measures and ROC curves. We will also seek evaluation of our work by submitting articles for peer review to journals in library and information science and other relevant disciplines. Presenting on our work, both in progress and at the conclusion of the grant period, will provide another opportunity to evaluate our work. See the timeline of activities below, where we have specified what kinds of evaluation and public outreach we plan over the three years of the grant period.

Project Design

Performance goals and outcomes

Our project objective is to develop a methodology and corresponding software that will enable librarians and archivists to utilize image analysis for identifying, isolating, and making more readily available specific kinds of content in large digital collections of historic newspapers. To this end, we will focus on developing image analysis procedures and software to identify poetic content and advertisements in historic newspapers over two centuries and continents, representing multiple languages and types of newspapers. We have identified the following performance goals and outcomes:

1. Analyze and verify the intelligent image analysis approach.
   a. Perform additional verification.
   b. Develop the system to be newspaper agnostic.
   c. Develop the system to be type agnostic.
   d. Develop the system to be language agnostic.
2. Scale and revise intelligent image analysis approach.
3. Distribute metadata and develop a new digital collection.
   a. Compile and distribute metadata.
   b. Develop an alpha version of a digital archive.
4. Disseminate results.

Activities

In order to meet our performance goals and objectives, we will pursue the following activities. We have correlated each activity below to a project goal/outcome above (e.g., "1.a." corresponds to "Analyze and verify the intelligent image analysis approach: perform additional verification.").

1.a. Extend the approach developed with our NEH Start-up Grant to analyze a minimum of 1 million additional newspaper pages from Chronicling America, which includes U.S. newspapers from the period 1836–1922. Analyze and verify results.

1.b. Extend the existing solution to identify poetic content in the Burney Collection of 17th and 18th century British newspapers, housed at the British Library and digitized by Gale/Cengage Learning. The Burney Collection material is similar enough to that in Chronicling America that it should prove as a further proof of concept of the broader applicability of Aida tools. Yet the Burney Collection's differences—in typography, layout, the paper on which the material was printed, the ways in which the collection was digitized, among others—will also prove a valuable testing ground for refining the existing solution.

1.c. Develop a new, second classifier for identifying advertisements. For both types of content (poetic and advertising), determine which balance of precision and recall are viable and appropriate: How much of the desired content do we need to find for the project to be successful? How much "noise" (non-desired content) is acceptable?

1.d. Analyze Western, non-English language newspapers from the Nebraska Newspapers project, including Czech-language papers, to further test the idea that an image-based approach should be language-agnostic. (Results reported in our *D-Lib* article include some Spanish-language papers.) If time allows, extend also to non-English newspapers from Oklahoma newspapers, including German, Cherokee, and Choctaw.

2. Revise and make scalable pre-processing approaches such as noise filtering and binarization, page segmentation and feature extraction techniques, and classification mechanisms to accommodate greater variety of newspapers involved. Further collect data to inform our solution for trading off precision and recall, to facilitate experimentation.

3.a. Compile and distribute metadata about poetic and advertising content in the collections, utilizing an approach that allows us to de-silo the data both from individual collections and our project work. Options for distributing the metadata are one of the many "smaller" research questions embedded in this project, and we have added a metadata librarian to the advisory board to help us consider possibilities. We will consider whether a linked data approach to metadata dissemination is appropriate and how that might be completed. Discuss possibilities for Gale and Chronicling America to incorporate poem metadata into their resources.

3.b. Develop an alpha version of a digital archive of the poems from the Burney Collection newspapers, including page images and transcriptions of the poems, as evidence of one type of further scholarship that Aida's methods enable. Gale/Cengage has already signaled their agreement for such use of their digital images (see attached letter).

4. Develop a series of open access reports documenting successes and challenges to image analysis of digitized collections. Distribute software and data. Present at professional conferences and offer training workshops so librarians and archivists can learn about the methods and how to use the software.

See the timeline of activities in the Project Resources section for a further delineation of activities.

Preliminary work and planning

Our current NEH Digital Humanities Start-up Grant, which concludes June 30, 2016, represents a significant area of preliminary work and planning. During our start-up project, we trained a classifier for identifying poetic content in historic newspapers, and we published early results of this work in *D-Lib*. We reported there training and testing statistics: 90.58% precision and 79.4% recall at the training stage, and 74.92% precision and 61.84% recall at the testing stage. Our article is available at dlib.org/dlib/july15/lorang/07lorang.html. This article includes a full description of our methods. In addition, we make available all of our interim reports on our project website, aida.unl.edu, along with news stories (such as one done by our affiliate NPR station in Lincoln, NE) and promotional materials prepared by the UNL Office of Research. In the last stage of our current grant work, we have been developing a use case and case study related to poetic content in Chronicling America newspapers from the period 1836–1840 (a total of about 25,000 pages). This case study has allowed us to further refine our classifier. We will submit a white paper to NEH in September with the results of this work.

Further, we have already gathered many poems manually. The UVA team has amassed on the order of 4,000 poems from the Burney Collection papers. In an earlier stage of work, members of the UNL team amassed a database of over 3,000 poems published in U.S. newspapers from 1830–1890—a drop in the bucket of the total number of poems published in papers during the period. This labor-intensive task has proven the underlying premise that there is a massive trove of undiscovered verse in the pages of historic newspapers and that we need computational approaches for locating and connecting users with these materials.

Project Resources: Personnel, Time, Budget

Personnel

This project is directed by Elizabeth Lorang, Leen-Kiat Soh, and John O'Brien and is a joint project of the University of Nebraska–Lincoln (UNL) and the University of Virginia (UVA). Lorang is Associate Professor of

Libraries at UNL and Faculty Fellow of the Center for Digital Research in the Humanities (CDRH; cdrh.unl.edu). Soh is Professor of Computer Science & Engineering at UNL. O'Brien is NEH Daniels Family Distinguished Teaching Professor in the English Department at UVA and associate fellow of the Institute for Advanced Technology in the Humanities (IATH; iath.virginia.edu). Institutional partners are CDRH and IATH.

With their individual and combined expertise, Lorang, Soh, and O'Brien are well qualified to successfully complete project tasks. Lorang has a master's degree in Library Science & Information Technology and a Ph.D. in English, with a focus on historic newspapers and poetry published in newspapers. She has been active in digital humanities and digital library work for more than a decade and has extensive experience managing a range of digital research projects, from expansive projects such as the Walt Whitman Archive (whitmanarchive.org) to smaller-scale projects such as the Aida start-up grant work and the digital scholarly edition "'Will not these days be by thy poets sung': Poems of the *Anglo-African* and *National Anti-Slavery Standard, 1863–1864*," published at scholarlyediting.org. For the proposed project, Lorang will be responsible for serving as project manager; contributing domain expertise from library and information science, particularly with regard to description, identification, and discovery; analyzing results of classification for Chronicling America images; and serving as the primary supervisor of a graduate research assistant enrolled in the Graduate Certificate Program in Digital Humanities at UNL. For a more specific breakdown of activities in which Lorang will be involved, see the activities timeline below.

Soh has conducted research in multiagent systems, intelligent systems, and image processing, with main applications in computer-aided education systems, multiagent simulations, and intelligent assistive systems. In image processing, Soh has built several tools, including ARKTOS, which classifies Arctic sea ice types based on analyzing satellite imagery. He has also applied data mining and machine learning techniques to a range of datasets and data types. Soh has published more than 150 papers on his research. Lorang and Soh have worked together since 2013 and have co-supervised four research students. They have also published together in *D-Lib, The Magazine of Digital Library Research* (July/August 2015). Soh will be responsible for contributing domain expertise in image processing and analysis, leading algorithm and software development, and serving as the primary supervisor of graduate research assistants from Computer Science & Engineering at UNL. For a more specific breakdown of activities in which Soh will be involved, see the activities timeline below.

O'Brien has published widely on eighteenth-century British literature and theater including two scholarly monographs: *Harlequin Britain: Pantomime and Entertainment, 1690–1760* (Johns Hopkins UP, 2004), and *Literature Incorporated: The Cultural Unconscious of the Business Corporation, 1650–1850* (University of Chicago Press, 2016). He is also the coeditor (with Brad Pasanek) of an online edition of Thomas Jefferson's *Notes on the State of Virginia* (http://jefferson-notes.herokuapp.com/). O'Brien will be responsible for testing Aida software with the Burney Collection of Eighteenth-Century Newspapers, working with the UNL team to refine the classifiers, and for building, with the help of programmer analysts at IATH, the alpha version of a digital collection of poetry amassed from the Burney archive. For a more specific breakdown of activities in which O'Brien will be involved, see the activities timeline below.

The project directors will be joined by 4 graduate student researchers. Three of the students will work at UNL and one at UVA. Two graduate research assistants (GRAs) at UNL will be students from Computer Science & Engineering (CSE) and will be responsible for implementing the project's conceptual ideas in the source code. In addition to the CSE GRAs, the UNL team will include a GRA in Digital Humanities (DH). With Lorang, this student will participate in digital library research, preparing materials for analysis, training classifiers, evaluating results, and performing additional human analysis. The GRA at UVa will be a graduate student in Digital Humanities, who will work throughout the year to prepare materials for analysis, analyze digitized pages from the Burney Collection, and help build the database of poems from the collection. Both UNL and UVA have sizable cohorts of graduate students who are developing expertise in digital methodologies and literary forms, as well as a long history of success with projects at the intersection of computational methods, newspapers, and poetry. GRAs will join other members of the project team as authors on conference presentations and publications.

IATH co-directors Worthy Martin and Daniel Pitti will serve in a consultative role. Programmer analysts at IATH will be responsible for constructing the alpha version of the database of poems culled via Aida software from the Burney Collection and of an interface for users to access them.

A server administrator at UNL will ensure maintenance and uptime of the development server on which the work will take place, including setting up a virtual machine for our project development and adding server hard drives and backup for storing page images, image snippets, processed images, and active research data.

Time

Project directors will each spend the time necessary to complete project objectives. Lorang has a 12-month appointment and research apportionment of 40%, which follows the UNL Libraries apportionment guidelines, and will spend as much of this research apportionment as is required to meet the grant objectives for the duration of the three-year project. In the first 4 months of the grant, Soh will be on research leave, and his research leave project is focused on Aida development. This research leave will contribute to overall objectives. Beyond his research leave, Soh will commit a minimum of one summer month to the project in year one and .5 summer months in both years two and three, as well as additional research time necessary to complete the project objectives throughout the academic year and summers. O’Brien has a twelve-month appointment for the 2016–2017 academic year and a nine-month appointment thereafter, with 50% of his appointment allocated to research. He will spend at least half of that 50% on this project during the academic year, and at least one month during each of the summers of this project.

GRAs at the University of Nebraska–Lincoln will work between 10 and 20 hours per week during the academic year, depending on their appointment, for each of the three years of the grant, based on their individual appointment. The graduate student at the University of Virginia will work an average of 11–12 hours/week for the duration of the grant period, including throughout the academic years and summers.

Timeline of activities

December 2016–May 2017

- Perform computational analysis of historic newspaper characteristics (Soh; CSE GRAs; Lorang)
- Reorganize existing code base to enable flexible experimentation and to streamline batch running processes (Soh; CSE GRAs)
- Publish code to GitHub repository (Soh; CSE GRAs)
- Develop GitHub pages version of project website linked to code repository on GitHub (Lorang)
- Prepare "ground truth" poetry datasets from Chronicling America and the Burney collection (Lorang; DH GRA, UNL; O'Brien; DH student, UVA)
- Convene monthly conference calls with advisory board (All)

June–November 2017

- Continue computational analysis of historic newspaper characteristics (Soh; CSE GRAs; Lorang)
- Hold project meeting and development sprint (All)
- 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)
- Analyze a minimum of 5,000 pages from the Burney Collection (O'Brien; DH student, UVA)
- Design, develop database of poems from Burney Collection (O’Brien; DH Student, UVA; IATH team)
- Convene monthly conference calls with advisory board (All)

December 2017–May 2018

- Prepare first open access report documenting success and challenges of year one work (All)
- Develop pre-processing approaches to accommodate a greater variety of newspapers (Soh; CSE GRAs)
- Present on work and/or lead a workshop at Code4Lib 2018 (Lorang; Soh; and/or DH GRA, UNL)
- Present on work and/or lead a workshop at ASECS 2018 (O’Brien; DH student, UVA)
- 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)
- Continue design, development phase of database of poems (O’Brien; DH student, UVA; IATH team)

- Convene monthly conference calls with advisory board (All)

June–November 2018

- Develop and test classifier for advertising content (Soh; CSE GRAs)
- Deploy pre-processing approaches and full processing pipeline for poetic content on previously processed Chronicling America pages; analyze and verify results (Lorang; DH GRA, UNL)
- Deploy pre-processing approaches and full processing pipeline for poetic content on previously processed Burney Collection pages; analyze and verify results (O'Brien; DH grad student, UVA)
- Hold project meeting and development sprint (All)
- Continue refining poetic content classifier as necessary (Soh; CSE GRAs)
- Lead a workshop at Digital Library Federation Forum to train participants on the software and get community feedback (UNL team)
- Investigate strategies for sharing metadata with originating collections and strategies for de-siloing the data (Lorang; O'Brien; DH GRA, UNL; DH student, UVA)
- Convene monthly conference calls with advisory board (All)

December 2018–May 2019

- Prepare second open access report documenting success and challenges of year two work (All)
- Refine classifiers and pre-processing approaches (Soh; CSE GRAs)
- Deploy pre-processing approaches and full processing pipeline for advertising content on previously processed Chronicling America pages; analyze and verify results (Lorang; DH GRA, UNL)
- Deploy pre-processing approaches and full processing pipeline for advertising content on previously processed Burney Collection pages; analyze and verify results (O'Brien; DH student, UVA)
- Process at least one million pages from Chronicling America (Lorang; DH GRA, UNL)
- Process entire Burney Collection (O'Brien; DH student, UVA)
- Generate/collect bibliographic information and other metadata for poems identified in Burney Collection (O'Brien; DH student, UVA)
- Generate/collect bibliographic information and other metadata for poems identified in Chronicling America (Lorang; DH GRA, UNL)
- Prepare "ground truth" datasets for poetic content and advertising content in languages other than English. Newspapers will include Czech-language newspapers from Nebraska newspapers and newspapers in at least one additional non-English language (Lorang; DH GRA, UNL)
- Present on work at MLA Annual Meeting (O'Brien; DH student, UVa)
- Convene monthly conference calls with advisory board (All)

June–November 2019

- Hold project meeting and development sprint (All)
- Complete database of poems from Burney Collection (O'Brien; DH Student, UVA; IATH team)
- Deploy full processing pipeline on non-English language newspapers (Lorang; DH GRA, UNL)
- Make software available (Soh; CSE GRAs)
- Participate in Chronicling America conference; lead workshop on software training (UNL team)
- Perform final analyses of project results (All)
- Prepare one or more open access reports, emphasizing sharing of metadata with original collections, application of approach to non-English language materials, and overall project success and challenges (All)
- Convene monthly conference calls with advisory board (All)

Budget

We request $462,318 in support of the above grant activities. The majority of project funds will support graduate students, portions of summer salary for Soh and O'Brien, programming and consulting services at UVA, and computing infrastructure at UNL, including the purchase of server hard drives for both primary and backup storage as well as computing services support from the Department of Computer Science & Engineering at UNL. We will utilize development servers supported by the CDRH and CSE at UNL. See the Budget and Budget Justification documents for further breakdown and discussion of these expenses.

There will be significant institutional investment beyond the funds requested from IMLS. For example, Lorang will spend as much of her research apportionment as is necessary to meet project objectives, though we have requested no grant funds to support her time. Soh and O'Brien likewise will make contributions beyond the time supported by grant funds. Lorang, Soh, and O'Brien also will seek undergraduate and graduate student support at their home institutions for additional student workers.

Readers will notice an increase in our overall budget from our preliminary submission. This increase is due to several factors. First, the F&A rate at UNL increased to 53.5% since our preliminary submission, and the F&A rate at UVA is 58%. In our preliminary proposal, we used 51% as the F&A rate for all budget calculations. Second, co-PI Soh was promoted to full professor, and his salary and fringe benefits increase as of July 1, slightly beyond the increase we had originally anticipated. We are also responsible for fees of $8,995 to the UNL Computer Science department, which were not accounted for in our original budget. These fees are mandatory and are charged in accordance with institutionally approved policies. Further, while we shifted the vast majority of the computer science work to UNL (we had a imagined a 50-50 split in the preliminary proposal), we added some professional programmer time and administrative time on the UVA subaward in order to meet project objectives. We also determined that the cost of hiring a graduate research assistant at UVA was cost prohibitive (approximately $40,000/year), so we adjusted the UVA budget to include a greater percentage of O'Brien's time—for which we are seeking grant funds—as well as an hourly graduate student worker. Finally, in response to readers' comments on our preliminary proposal to emphasize dissemination to library and archival communities, we have increased our travel budget significantly from the preliminary proposal. This increase in travel funds is designed to allow members of the project team to travel to relevant professional conferences and offer workshops in both the larger methods as well as the specific software.

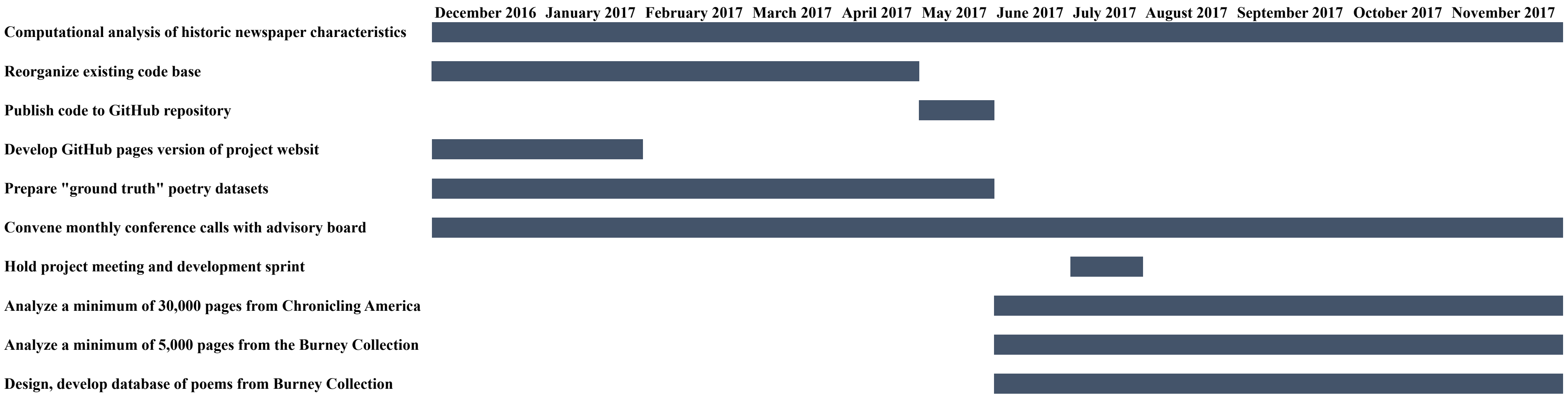
Project finances will be administered by the UNL University Libraries' grants administrator in cooperation with the UNL Office of Sponsored Programs and in compliance with institutional rules and state and federal law. Financial reporting will be handled by the UNL Office of Sponsored Programs. Subaward finances will be administered by the Office of Sponsored Programs at the University of Virginia, who will submit reports to the Office of Sponsored Programs at UNL.

Communications Plan

Our primary audience is librarians and archivists who create, curate, and manage large digital collections. We anticipate a secondary audience of researchers who use large digital collections. To reach these audiences, we plan to have a consistent public presence on the web at aida.unl.edu. We will migrate our existing website to the GitHub Pages platform, as we are using a public GitHub repository to make our source code available. The project website will serve as a portal to the code repository, to data sets deposited in the UNL Data Repository and elsewhere, to our series of open access reports disseminated via the UNL and UVA institutional repositories, and to all other grant products. The code repository will include appropriate technical documentation to allow others to use the software as well as to extend it. In addition, the code repository will provide a mechanism for users of the source code to create records of issues as well as to request features.

Lorang and Pitti will take the lead on outreach, promotion and dissemination to the libraries and archives communities. Lorang will coordinate promotion and outreach efforts, and all team members will promote the project to and seek feedback from researchers who use large digital collections. We plan to demonstrate the project through talks and workshops at the annual Chronicling America meeting, Code4Lib, the Digital Library Federation Forum, the Modern Language Association annual meeting, and the annual meeting of the American Society for Eighteenth-Century Studies, among others. These demonstration and training opportunities will be designed to facilitate audience engagement and involvement and will inform subsequent developments of our methods and software. We will also publish open access reports targeted at multiple audiences: at librarians interested in applying such a tool to large digital collections, either in English or in other languages; at computer scientists interested in the development of intelligent image processing; in scholars interested in text mining, network analysis, popular culture, and poetry.

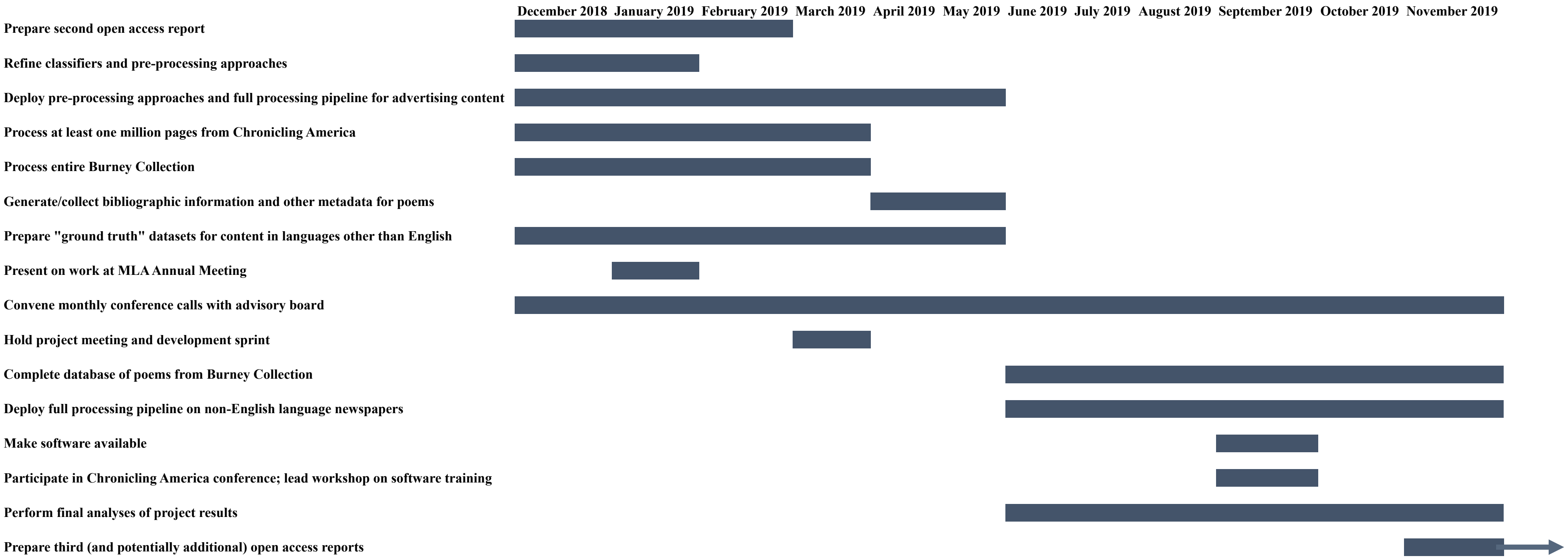
Schedule of Completion, Year 1



Schedule of Completion, Year 2

	December 2017	January 2018	February 2018	March 2018	April 2018	May 2018	June 2018	July 2018	August 2018	September 2018	October 2018	November 2018
Prepare first open access report documenting success and challenges of year one work	■	■	■									
Develop pre-processing approaches to accommodate a greater variety of newspapers	■	■	■	■	■	■						
Present on work and/or lead a workshop at Code4Lib 2018			■									
Present on work and/or lead a workshop at ASECS 2018				■								
Prepare "ground truth" datasets for advertisements	■	■	■	■	■	■						
Continue design, development phase of database of poems	■	■	■	■	■	■	■	■	■	■	■	■
Convene monthly conference calls with advisory board	■	■	■	■	■	■	■	■	■	■	■	■
Develop and test classifier for advertising content							■	■	■	■	■	■
Deploy pre-processing approaches and full processing pipeline for poetic content							■	■	■	■	■	■
Hold project meeting and development sprint								■				
Continue refining poetic content classifier as necessary	■	■	■	■	■	■	■	■	■	■	■	■
Lead a workshop at Digital Library Federation Forum												■
Investigate strategies for sharing metadata with originating collections							■	■	■	■	■	■

Schedule of Completion, Year 3



DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions
If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project (Check all that apply):

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
☑	Digital content	Part II
☑	Software (systems, tools, apps, etc.)	Part III
☑	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

The University of Nebraska Board of Regents Policies, containing amendments through December 2015, state that individual faculty members hold the rights to University supported works that "do not involve use of substantial University resources," and the "University does not regard the provision of an employee's salary, office, usual library resources, usual facilities and equipment, and office staff, or personal computers as constituting 'substantial use of University Resources.'" At this stage, we believe that PIs Lorang and Soh will hold intellectual property rights to content, software, and datasets we create, excluding contributions by the UVA team. Depending on the nature of contributions from students, it remains possible that intellectual property rights for certain creations will be held by the University of Nebraska.

Project created software will be issued under a GNU General Public License (GPL), v. 3. GPL is a free software license that allows users to run, share, and modify the software. We will, in part, be building on machine learning software licensed under these terms, and we seek to facilitate further development of our software on these terms.

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

At this stage, we do not believe that the University of Nebraska will assert any ownership rights for grant products. Grant products will be made freely available with no conditions of use beyond those carried by the GNU GPL License and the Creative Commons license. Digital content published on official University of Virginia web sites is the intellectual property of UVA (see <http://www.virginia.edu/siteinfo/copyright>). All content created for this project will be made freely available to the public. We will not impose any conditions of use on published data.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

We will not create any content or products which may involve privacy concerns. We have obtained permission and cooperation from Gale for the use of their materials (see attached letter). We will not create any content or products that raise any cultural sensitivities.

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

The project team will collect digital content from Chronicling America and the Burney Collection newspapers, the latter digitized by Gale. We will process at least 1 million page images from Chronicling America and the images from the Burney Collection.

The majority of content created will be datasets and written reports. In addition, the University of Virginia team will create an alpha version of a database of poetic content identified in the Burney Collection, which will contain the first lines of at least 3,000 newly discovered poems and descriptive metadata about both the poems and the newspapers where they are located. This content will be available to the public via a web interface. Textual content will use widely used open source standards (e.g., HTML, CSS, and TEI XML). The project will use JSON, Jquery, Ruby, Rails, Sinatra, and Passenger for web-based data queries to a PostgreSQL database.

The remainder of this section documents our digital stewardship plan for the written reports and for the database of poetic content. Digital stewardship of software and datasets is described in Part III and Part IV below.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

The image feature datasets and derivate images will be created on a Linux development server maintained by the Center for Digital Research in the Humanities at UNL. This server is physically located on the UNL campus and is administered by CDRH staff. Project team members will have direct access to the server for running software and analyzing results. The datasets and derivative images will be created by Java and Python-based software. The database of poems created by the UVA team will be stored on one of IATH's four servers, which are located at the University of Virginia's Information Technology Services (ITS) data center. Currently, the project is on a Dell PowerEdge R720 server with 128 GB RAM and 4TB of internal storage space. The Operating System (OS) on this server is CentOS 6 (64 bit), an Enterprise class Linux OS.

The open access reports will be developed on laptop and desktop computers using standard word processing software by members of the project team.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

Images of poetic content will be obtained as JPEG200s from Chronicling America and as TIFFs from the Burney Collection. For the purposes of image processing, these will be used to creative JPEG derivatives. The prototype database developed by the UVA team will deliver access JPEGs of poetic content at a minimum resolution of 72dpi. Textual content will be stored as XML and HTML.

Open access reports will be developed using word processing software, such as Microsoft Word, Libre Office, or Google Docs, and will be exported as PDF documents.

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

Quality control will be exercised via an iterative development process between the UNL and UVA teams. Development of the poetry database will be carried out primarily at UVA, but UNL team members also have extensive experience in developing content collections and accompanying standards. UVA PI O'Brien will work closely with the student research assistant on quality control through the grant period. We will begin testing with an already identified set of 1,000 poems, then expand to include new poetic material that is identified by the Aida team's software.

The technical quality of the open access reports will be evaluated by the Scholarly Communications team at the University of Nebraska-Lincoln Libraries prior to submission to the UNL institutional repository, Digital Commons.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

The UVA team will utilize IATH servers to host all of the IATH's publicly accessible projects, using dedicated virtual servers that ensure that each project has resources sufficient to provide efficient and reliable processing and response. ITS provides 24x7x365 maintenance of the servers as well as backup and security services. IATH's system administrator also maintains redundant copies of IATH server data on a set of external hard drives. This level of support is provided indefinitely for all IATH projects and will enable the prototype to remain accessible and stable (in the state it has reached by the end of the grant period) for the foreseeable future. IATH has a strong track record of maintaining its projects (some have been running for twenty years) and providing institutional support that includes technical support to maintain access and functionality of the project's element, as well as development advice and help.

Open access reports will be submitted to Digital Commons, the UNL institutional repository. The institutional repository is hosted by bepress, which has "a multi-tiered disaster recovery plan utilizing fail-over servers and regular on-site and off-site backups" and supports LOCKSS, an Open Archival Information System-compliant preservation solution. The University Libraries will be responsible for the long-term management and maintenance of these assets.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

The UVA team will store descriptive and preservation metadata about poetic content in a database. The metadata will be available to users to be downloaded as open data. The UVA team will use MODS (http://www.loc.gov/standards/mods/) for descriptive metadata for each identified poem and Premis (http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf) for metadata essential to the long-term preservation of the files representing each poem.

Report metadata will be created by the project team at the time of deposit in the UNL institutional repository, following UNL institutional repository standards and metadata structures.

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

Metadata files will be preserved with the same methods as described in section B.2.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

Because the poem database being developed is a prototype, the project team will not be making contributions to DPLA, though we will consult with the metadata coordinator at DPLA and UVA Library specialists who coordinate UVA contributions to DPLA. We will also explore exposing the MODS descriptive data as RDF and a SPARQL endpoint (using Neo4J as the graph database).

UNL's institutional repository software, bepress, is widely discoverable on the open web, including in Google Scholar and via searches of the standard web.

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

The poem database will be made publicly available when it reaches a point of technical and intellectual stability, as accepted by the project staff. It will be distributed via the internet with a public interface that allows users to browse and search through content.

Reports will be openly available online via the UNL institutional repository.

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

CDRH's main page is cdrh.unl.edu and contains links to current and past projects. CDRH has developed expertise in metadata projects as well as in developing thematic and content-based digital collections. IATH's main page is iath.virginia.edu and contains links to our current and past projects. IATH projects all have digital contents and contain some combination of text, 2D and 3D images, audio, and maps.

The UNL Institutional Repository is online at http://digitalcommons.unl.edu/

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

Our software retrieves and processes large collections of image data sets, extracts features from the images, computes feature values for the features, and trains a machine learning classifier to identify specific types of content. The software uses a neural network to analyze and classify images. The primary audiences of this software are archivists, librarians, and others creating digital collections who are in need of additional tools to help in extracting, in an automated way, informational value out of large digital collections, and researchers using these collections.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

There is an increasing array of computer vision software including OpenCV, Pastec (based on OpenCV), and Google's Cloud Vision API and related DeepDream software. Industry applications of these and related software include Google Goggles, Amazon Firefly, and Flickr's Magic View. These software principally function on images of artwork and contemporary mass photography. No existing software that we have found focuses on computer vision for textual materials in the type of application we propose.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

Our primary programming languages for software are Java for the main source code base (image processing and analysis) and Python for the batch operationalization code base (retrieving images from Chronicling America and processing them in batches). Java is a rich, high-level language, is platform-independent, with vast community support, and is versatile with its large sets of libraries. Furthermore, the machine learning and data mining programs from the open source Weka repository are also written in Java, thus supporting our use of Java as the main programming language. Python, a scripting language, allows for more efficient data handling especially with large sets of files, which is necessary in the initial stages of setting up our image data. Python also is a well-supported language with a wide community and support base.

B.2 Describe how the intended software will extend or interoperate with other existing software.

This software will extend applications of the machine learning software developed for the Weka repository (described below). Our software can be coupled with other programs using bash scripting (to add additional layers of the processing pipeline, for example).

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

We make use of the open source data mining and machine learning programs (also Java-based) available online on the Weka website (http://www.cs.waikato.ac.nz/ml/weka/), a popular and highly regarded repository by computer scientists especially in the area of artificial intelligence and machine learning.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

We will use a combination of pseudocode, algorithm description, inline documentation, readme documentation, and a user guide for maintaining and updating technical documentation for users of the software. This holistic approach will provide interested users with the conceptual and methodological underpinnings of our software as well as practical technical details for using the software, including system requirements. In this way, we will reach different types of users: those who want to use the software as is, and those who want to extend and build upon it. That is, the documentation will reflect both how the software works--including the values and assumptions embedded in the software--as well as how to use it. Documentation will be folded into overall project development and will be reviewed at each significant iteration of the project.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

Software systems and tools developed by the Center for Digital Research in the Humanities are available at https://github.com/CDRH. Aida code will be included here.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

As stated in Part 1, we do not believe that our organization will assert ownership rights over the software created. We will need to review this assessment annually, with the key variable being the nature of the contributions of the student researchers. At this time, we anticipate the students will implement the ideas of faculty leads, but if the students make substantive intellectual contributions, we will need to discuss ownership with the University of Nebraska.

We will make the software under a GNU General Public License (GPL), v. 3. GPL is a free software license that allows users to run, share, and modify the software. We will not impose restrictions on the access and use of the product.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

We will make the source code available via a public GitHub repository. We will link to this repository from our project website as well as from publications emerging from the project.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository: GitHub
URL: https://github.com/CDRH

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

Datasets created for this project include: data sets of extracted feature values and datasets documenting where poetic content and advertising content appear in the digitized newspaper collections. These datasets will be created each time we run the source code and will be generated by the software. The data will be used for the purposes of training machine learning classifiers and ultimately for classifying pages as containing or not containing particular types of generic content.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

Our proposed data collection or research activity does not require approval by an institutional review board.

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

We will not collect any personally identifiable information, confidential information, or proprietary information.

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Not applicable.

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Our datasets of feature values will be generated by our software program, which will output alphanumeric data to a plain-text file format, such as .txt and .csv. We will include a data dictionary that identifies the data fields and the algorithms used to compute the different feature values.
Datasets of where poetic content and advertising content appear in the historic newspapers will be created using the feature value datasets and image classification values. These datasets may require manual manipulation by project staff. While we may ultimately make this data available in a number of formats, we will commit to maintaining plain-text versions of the data made available in .txt and/or .csv formats.

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

For each instance in which we run the software and create datasets, we will create accompanying documentation that documents the input data, version of the software used (software will be versioned using git and made available in a GitHub repository), the environment in which the code was run, and any variables to normal procedures outlined in project documentation. We will create persistent identifiers for the datasets and cue documentation to those persistent identifiers. This information will be available in a plain-text file along with the dataset it describes.

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

We will deposit the data in the UNL Data Repository. We also may pursue deposit of the data in additional repositories to increase findability of the data. The Data Repository is managed by the University of Nebraska-Lincoln Libraries and Information Services and is "a secure site for storage of data collections that are no longer actively in use [allowing] the researcher to stably retain data for future use and/or sharing with other interested parties." The University Libraries will be responsible for long-term data management and maintenance. We will maintain the data sets for a minimum of five years following the completion of the project. We will link to repository data from our project website and from publications.

8. Identify where you will be publicly depositing dataset(s):

Name of repository: University of Nebraska-Lincoln Data Repository
URL: https://dataregistry.unl.edu/

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

We will review this data management plan annually with the data curation librarian in the UNL University Libraries. PI Lorang will be responsible for conducting this annual review with the UNL data curation librarian and monitoring implementation with the data curation librarian and the project team. We will provide information and updates on data management in annual reports.

Original Preliminary Proposal

Extending Intelligent Computational Image Analysis for Archival Discovery

Project Summary: This proposal seeks research grant funding in the estimated amount of $403,000 for researchers at the University of Nebraska-Lincoln and the University of Virginia--with expertise in information science, computer science, and literary studies--to investigate the use of image analysis as a methodology for content description, discovery, and information retrieval in digitized collections of historic materials. Our novel approach leverages machine learning to build an intelligent computational system that first extracts visual cues from image training sets, then trains a classifier to recognize these visual cues, and subsequently deploys the system to process new image snippets. The Image Analysis for Archival Discovery team (Aida; aida.unl.edu) began this work with an NEH Office of Digital Humanites Start-up Grant, and we have shared preliminary results in "Developing an Image-Based Classifier for Detecting Poetic Content in Historic Newspaper Collections" (*D-Lib Magazine*, July/August 2015, http://dlib.org/dlib/july15/lorang/07lorang.html). We seek now to extend this work across a more diverse range of digitized newspapers and textual forms and to assess the broader potential of image analysis as a methodology for information identification, discovery, and retrieval in massive digital libraries. In particular, we propose to continue our ongoing case study to identify poetic content in digitized historic newspapers; demonstrate the potential research outcomes when we are able to identify new content in digitized collections at scale; develop new machine-learning classifiers for additional types of content in historic newspapers; study alternative machine-learning methods to those we have thus far used; and significantly extend the scholarly literature on these topics.

Directors and partners: This project is directed by Elizabeth Lorang, Leen-Kiat Soh, and John O'Brien and is a joint project of the University of Nebraska-Lincoln (UNL) and the University of Virginia (UVA). Lorang is Associate Professor of Libraries at UNL and Faculty Fellow of the Center for Digital Research in the Humanities (CDRH; cdrh.unl.edu). Soh is Associate Professor of Computer Science & Engineering at UNL. O'Brien is NEH Daniels Family Distinguished Teaching Professor in the English Department at UVA. Institutional partners are CDRH and the Institute for Advanced Technology in the Humanities at the University of Virginia (iath.virginia.edu).

Relevance to agency priorities and fieldwide need: Our work addresses the challenges of description, analysis, and discovery within digitized collections. This project falls under the IMLS "National Digital Platform" priority. In particular, the goal of this project is to improve the discoverability of digital content and to do so by providing approaches for identifying content at scale. Our project also engages the IMLS question "What will move library and archival services in the United States forward?" When fully realized, image processing approaches and intelligent classifiers may become part of the standard toolkit of archivists creating digital collections and provide a mechanism for more fine-grained description than is currently feasible. In another scenario, researchers may use our classifiers and approach to find items of relevance in digital collections, if they have access to large sets of digital images. For the present project, we remain focused on digitized newspapers, but at a later stage we plan to work other materials as well.

Impact: This methodology stands to have significant impact in thinking about and imagining new modes of access and discovery in digital collections. Images often have more information potential than the limited metadata associated with them, and to leverage this full information potential, we must expand our methods of analysis to include digital images. Further, developing image processing and machine learning technologies for identification and discovery, such as for visually distinctive forms and genres, means that we can deal with multi-language corpora for certain types of research questions, since our intelligent image analysis need not understand the text of the material.

Performance Goals and Outcomes

1. ***Analyze and verify the intelligent image analysis approach.***
    a. *Additional verification:* Extend the approach developed with our NEH start-up grant to analyze a minimum of 1 million additional newspaper pages from Chronicling America, which includes U.S. newspapers from the period 1836-1922. Analyze and verify results.
    b. *Newspaper agnostic:* Extend the existing solution to identify poetic content in the Burney Collection of 17th and 18th century British newspapers, housed at the British Library and digitized by Gale/Cengage Learning. The Burney Collection material is similar enough to that in Chronicling America that it should prove as a further proof of concept of the broader applicability of Aida tools. Yet the Burney Collection's differences--in typography, in layout, in the kinds of paper on which the material was printed, in the ways in which the collection was digitized, among others--will also prove a valuable testing ground for refining the existing solution.
    c. *Type agnostic:* Develop a new, second classifier for identifying an additional type of content--such as market or weather data, sports scores, advertisements, or other visually distinctive *textual* newspaper content. For both types of content, determine which balance of precision and recall are viable and appropriate: How much of the desired content do we need to find for the project to be successful? How much "noise" (non-desired content) is acceptable?
    d. *Language agnostic:* Analyze Western, non-English language newspapers from the Nebraska Newspapers project, including Czech-language papers, to further test the idea that an image-based approach should be language-agnostic.
2. ***Scale and revise intelligent image analysis approach.*** Given the analysis and verification results of Objective 1, revise and make scalable pre-processing approaches such as noise filtering and binarization, page segmentation and feature extraction techniques, and classification mechanisms to accommodate greater variety of newspapers involved. Further collect data to inform solution on trading off precision and recall, to facilitate experimentation.
3. ***Develop and meta-tag digital archives.*** Develop an alpha version of a digital archive of the poems from the Burney Collection papers, including page images and transcriptions of the poems, as evidence of one type of further scholarship that Aida facilitates. Gale/Cengage has already signaled their agreement for such use of their digital images. Discuss possibilities for Gale to incorporate poem metadata into their resource. Provide metadata based on page images and transcriptions of the poems for meta-tagging the Chronicling America digital archive newspapers that we have processed.
4. ***Disseminate results.*** Develop a series of open access reports documenting the successes and challenges to image analysis of digitized historical newspapers. Make software and results available.

Work Plan: We will undertake this work over a period of three years, and the project team will include Lorang, Soh, and O'Brien, as well as two graduate research assistants at each institution for two of the three years. Team members will work collaboratively on all objectives, with the UNL-based team taking the lead on Chronicling America and Nebraska Newspapers work and the UVa team taking the lead on Burney Collection work. For administrative purposes, UNL will be the lead institution, with UVA acting as subcontractor.

Budget: We anticipate the total cost at this time to be $403,000. Not included in this cost is the research time of Professors Lorang and O'Brien, who will use the necessary amount of their research apportionment in order to complete the project. Our estimated budget breakdown is: $144,000 for graduate research assistant (GRA) salaries; $58,464 GRA tuition remission; $13,708 GRA health insurance; $28,195 summer salary for Soh; $7,895 benefits for Soh; $10,000 hardware, computer services; $9,500 travel, including 4 team members' travel to meeting and project sprint at UVA, 3 team members' and IATH member's travel to meeting and project sprint at UNL; and a final meeting of 2 team members' travel to UNL or UVA; and $131,238 in F&A, calculated at 51% of modified total direct costs.