Programmatic Extraction of ‘Documents' from Web Archives

The UNT Libraries and the UNT Computer Science and Engineering Department are seeking IMLS support under the National Digital Platform category for a two year research project to evaluate the use of machine learning algorithms to successfully identify and extract publications contained in existing Web archives. Identifying these documents will empower libraries, archives, and museums to meet their curatorial missions.

As Web archives continue to grow in the United States and around the world, the tools and services to interact with large Web archives need to continue to improve. One area for new tools is in extracting parts of Web archives in order to build different kinds of collections for users. With this kind of collection building, institutional repositories could extract articles and publications from faculty websites that are contained in Web archives of a university's’ domain. State libraries and archives could extract documents published by agencies throughout their states and add them to statewide collections of publications. Federal government information specialists would be able to identify content-rich publications from Web archives of the federal .gov domain to include in their local collections. All of these use cases are too time consuming with existing workflows and tools, but with methods and workflows identified during this research project, the process would become easier.

We will complete a two-track effort during the proposed project. First, using observation and structured interviews, we will collect data from Web archive curators and content selectors in institutional repositories, state and local government collections, and federal Web archive collections. This qualitative data gathering will inform the second track of work, which will focus on the training of machine learning models to aid in the automatic extraction of documents of interest to collection curators. The project team includes PI Mark Phillips, Associate Dean for Digital Libraries at the UNT Libraries, Co-PI Dr. Cornelia Caragea, Assistant Professor in the UNT Computer Science and Engineering Department, and Lauren Ko, Head of the Software Development Team in the Digital Libraries Division of the UNT Libraries, along with two graduate research assistants. To help guide the research, the project will convene an advisory board of collection professionals and machine learning researchers. This project will provide insight into how collectors select resources from the Web for inclusion in their collections, create open datasets that can be used in the training of new machine learning models, and disseminate white papers and publications detailing the project findings.

As institutions continue to adopt Web archiving technologies to harvest and preserve born-digital publications from the Web it will be increasingly important to find better ways of providing access to these resources. Incorporating algorithms and processes from the field of machine learning into the process of building locally relevant and meaningful collections for users offers a promising step in this direction. This project will investigate the ability and practicality of leveraging machine learning techniques and applications as part of the collection management processes for large corpora of digital content in order to extract documents from Web archives and create meaningful digital collections for the end user.

1. Statement of Need

Increasingly available and accessible Web publishing technology necessitates that many cultural heritage organizations, including libraries, archives, and museums collect materials from the Web with Web archiving technologies. These organizations need new capabilities for identifying and selecting materials in accordance with their respective missions and collection policy scopes. At present, selecting relevant content from Web archives is a daunting endeavor, in large part because most Web archives harvest and store web-published materials in a manner optimized for preservation and not in a manner that supports information discovery. Thus, locating usable information is truly an act of finding the proverbial needle in the haystack. Identifying high value documents published as PDF and Microsoft Word files in a large amount of content in Web archives holds promise as a solution that will enable extension of collection development practices to this new class of materials. The audience for this project is any library, archive, or museum intent on mining Web archives to extract unique and high value documents for their local collections.

1.1 Web Archives

A growing number of research libraries, museums and archives around the world are embracing Web archiving as a mechanism to collect born-digital material made available via the Web. Between the membership of the International Internet Preservation Consortium which has 55 member institutions (International Internet Preservation Consortium, 2017), and the Internet Archive's Archive-It Web archiving platform with its 529 collecting organizations (Archive-It, 2017), there are well over 584 institutions currently engaged in building collections with Web archiving tools. The amount of data that these Web archiving initiatives generate is typically at levels that dwarf traditional digital library collections. As an example, in a recent impromptu analysis, Jefferson Bailey of the Internet Archive noted that there were 1.6 Billion PDF files in the Global Wayback Machine (Bailey, 2017). If just 1% of these PDFs are of interest for collection organizations, that would result in a collection larger than the 15 million volumes in HathiTrust (HathiTrust, n.d.).

Over the last few years, an increasing number of institutions in the United States have become involved in Web archiving activities. This is evidenced by the results of a 2016 survey conducted by the National Digital Stewardship Alliance (2017), which surveyed organizations in the U.S that were actively involved in, or planning to start, programs to archive content from the Web. The 103 respondents came from a broad range of cultural heritage institutions, such as universities, government organizations, museums, and archives.

The study found that 79% of the surveyed institutions had Web archiving programs that were currently in production. Additionally, 5% of institutions were actively testing Web archiving programs and another 15% were planning on pursuing a Web archiving program in the near future. In total, a full 98% of respondents were actively planning on archiving Web content (National Digital Stewardship Alliance, 2017, p. 8).

1.2 Machine Learning in Libraries

In the past few years, other research and demonstration grants have been funded by the Institute of Museum and Library Services (IMLS) and other granting organizations in the area of utilizing machine learning to interact and create new understanding of patterns and organizations inherent to large collections of digital content. Among projects that use digital content to mine and extract new meaning and organizational patterns inherent in the resources themselves, was a research project (LG-06-08-0057-08) to explore the use of topic modeling as a way of improving search and discovery in HathiTrust (Hagendorn, Kargela, Noh, and Newman, 2011). More recently, the project "Improving Access to Time-Based Media through Crowdsourcing and Machine Learning" (LG-71-15-0208-15) incorporated machine-learning technologies related to machine transcription to collections

of audio recordings. Another project, again working with the HathiTrust corpus, this time with the University of Illinois and University of Notre Dame with support from the Andrew W. Mellon Foundation, in the "Use of Scale in Literary Study" project (Use of Scale in Literary Study, n.d.) investigated the use of new methods including machine learning to classify works of fiction from the HathiTrust collection into the different genres of fiction. These funded research projects demonstrate the potential for combining machine learning algorithms with data and content collected in libraries, archives, and museums. It is important that we investigate new ways of using our existing collections of content to help us better serve our users. The use of machine learning algorithms and Web archives is a prime example of leveraging our data and new tools to help provide increased access and use to our collections.

The availability and utility of general purpose machine-learning technologies and applications have allowed researchers, collectors, and curators working with large amounts of textual data and large digital files to perform data mining analyses and glean insights across individual works that were previously unobtainable. Could these same types of technologies and applications be used for purposes of creating and managing digital collections?

To these ends, project team members at the University of North Texas (UNT) will investigate the ability and practicality of leveraging machine learning techniques and applications as part of the collection management processes for large corpora of digital content in order to extract documents from Web archives and create meaningful digital collections for the end user. In particular, this project will research if these machine learning techniques and tools have reached a state of maturity that allows institutions to adapt and extract locally meaningful documents from existing Web archives and integrate them into digital library collections.

1.3 What Is a Document?

The words *document* and *publication* have a variety of meanings in different communities. Throughout this proposal, they are used interchangeably to denote a general class of content that has been created and distributed with the goal of communicating to an audience and disseminating scientific, cultural, or governmental information. This broad definition purposefully includes but is not limited to, external publications from agencies, articles, white papers, reports, and notices. The term *resource* is used more generically to describe content held within a Web archive that features characteristics such as multiple pages of content in a file format such as PDF or Microsoft Word. These resources can be investigated to see if they are documents that a cultural heritage organization would be interested in acquiring. The terms should ideally not cause issue as they are working terms to describe content of possible interest for collection (document/publication) and a superset of content with a particular set of characteristics (resource).

1.4 Use Cases for Extracting Documents

The value of the research proposed here is best communicated through three short use cases that demonstrate the broad applicability and great need of this project. Web archives contain a huge amount of content and are collected by a wide range of institutions to meet a variety of collection needs. The three use cases listed below will be addressed in this project.

1.5 Institutional Repository Use Case

According to the Cybermetrics Lab in Spain there are 391 institutional repositories (IR) operated by institutions across the United States (Cybermetrics Lab, 2017). These IRs are generally responsible for the identification and curation of the published faculty output from their institutions. This published output typically includes research and scholarly articles, white papers, conference posters, and presentation slides. Universities and colleges are a primary operator of IRs around the world and this holds true here in the United States. The hosted

Web archiving service Archive-It Web identifies 257 collecting organizations that are classified as "Colleges and Universities" (Archive-It, 2017). If there were a way for these institutions to further leverage the Web archives that they are collecting with services like Archive-It to help build these collections of faculty publications, it would be a huge benefit to a wide range of organizations across the country.

1.6 State Library and Archives Use Case

State libraries and state archives are a group of institutions that are often charged with collecting the published output of their state governments. Examples and approaches to this task differ across the country. In the past decade, the majority of these programs have moved from a print workflow, to a mixed print and born-digital workflow, to now a primarily born-digital focused workflow for identifying and collecting state publications. In many of these states they have moved from identifying specific publications from agencies or agencies' websites to a model where they are using Web archiving tools to harvest the websites of agencies on an ongoing basis. An example of this is the Texas Records and Information Locator (TRAIL) service in Texas (Texas State Library and Archives Commission, 2014) that uses the Archive-It service to collect Texas state government websites and provides access to these websites and the publications within the websites through the Archive-It search interface. There are currently 30 state library and archive collecting organizations that subscribe to the Archive-It service (Archive-It, 2017). If there were an automated way to identify publications within these collected Web archives, it could allow these state publications collections to provide item level access to publications that mirror the pre-Web archiving methods of providing access. Item level description and access in a traditional digital library platform would allow more granular descriptive metadata that would improve access to these resources.

1.7 Federal Publications Use Case

Web publishing has spread into most aspects of communication. Perhaps one of the greatest examples of this is the use of the Web as a platform for the United States federal government to inform its citizens about the work carried out on their behalf through federally created, funded, and sponsored work. Initiatives such as the End of Term Web Archives in 2008, 2012, and 2016 (http://eotarchive.cdlib.org/) worked to collaboratively harvest the transition between administrations of Presidents Bush, Obama, and Trump. These collections contain hundreds of millions of URIs, and tens of millions of PDF files. Archived files like PDF and Word Documents provide a rich potential resource to build collections if it is possible to effectively identify and suggest which publications might be the most likely to be in scope with a specific collection and its collection plan. A concrete example of this potential type of work is to identify technical reports funded by the federal government for inclusion in an initiative such as the Technical Reports Archive and Image Library (http://technicalreports.org/).

1.8 Summary

As libraries, museums, and archives continue to adopt Web archiving technologies to harvest and preserve born-digital publications from the Web, it will be increasingly important to find better ways of providing access to these resources. Combining algorithms and processes from the field of machine learning with the process of building locally relevant and meaningful collections for users offers a promising step in this direction. This project will investigate the ability and practicality of leveraging machine learning techniques and applications as part of the collection management processes for large corpora of digital content in order to extract documents from Web archives and create meaningful digital collections for the end user.

2. Project Design

This project will bring together researchers from two disciplines with complementary areas of expertise: library and information science expertise in Web archives, digital libraries, and born-digital government information (Phillips) and computer science expertise in natural language processing, information extraction, and machine learning (Caragea). The multidisciplinary team is an important design feature of this project and will be crucial to its overall success.

The collaborative project is designed as a single project with two interlinked research tracks. The **Knowledge-Gathering Work Stream** is responsible for qualitative and ethnographic-based research methods, including observation, interviews, and analysis of existing collection policies that will help in designing features for the **Machine Learning Work Stream**. The Machine Learning Work Stream will use aspects derived from the knowledge-gathering work stream to train and evaluate machine learning models, workflows, and the quality of designed document features.

2.1 Goal, Methods, Assumptions, and Risks

Project Goal: The overarching goal of this project is to understand if machine-learning models can successfully identify content-rich documents from Web archives that align with various organizations' collecting plans. This goal will be achieved by increasing our understanding of the workflows, practices, and selection criteria of publication selectors through ethnographic-based observation and interviews. This increased understanding will inform the use of novel machine learning techniques to identify content-rich publications from PDF and Word documents collected in existing Web archives. This project has the potential to identify large quantities of content rich documents and publications that will need further curation activities, including selection, metadata creation, indexing, and discovery. These activities are important, but they fall outside of the scope of this project. However, they present important future work that would result from this successful project.

The project is designed around the following primary research questions:

1. *What characteristics and patterns do professionals tasked with building collection documents gathered from the Web use to identify these publications?*
2. *Can these characteristics and patterns be translated into generic features that can be used to accurately predict if a Web document should be recommended for a collection?*
3. *Can the machine learning models be reliably used on Web archive content to identify candidate documents for inclusion in collections, and how well do our models generalize to new (unseen) data?*

Research Methodologies: The project will make use of research methodologies from the domain of library and information science with an ethnographic center combined with machine learning techniques from computer science to answer the above research questions. The mixed-method approach of this project aims to produce a higher-quality final result by employing multiple data collection, analysis, and testing methodologies.
Project Risks: The main risk for this project revolves around identifying a suitable number of practitioners for observation and interviews. The inclusion of an advisory board described below will broaden our project team's network and will mitigate this potential risk. Another potential risk to the project is gaining access to representative Web archives for experiments. To combat this potential risk, the project plans to make use of existing extensive Web archives collected by the UNT Libraries over the past decade, which will ensure access to adequate data for use during the project.

Assumptions: This project has assumptions that we feel are reasonable for a research project of this type. First, we feel that there is a need to apply ethnographic-based research methods to this problem area because not

enough is known about the mechanics of selecting web-published material. The choice of observation and interviews for data collection over survey based data collection aligns with this assumption. Second, we assume that it will be possible to convert learned human practice such as the selection of born-digital publications for a collection into features that can be used to build successful machine learning models.

2.2 Activity Overview

Work Area 1: Knowledge Gathering Identify and work with practitioners who have a primary duty of selecting born-digital publications from the Web to add to existing collections of document resources.

Tasks

- Identify participants for observation and interviews using existing contacts and suggestions from the project's advisory board.
- Engage in qualitative data collection techniques with identified participants to conduct field observations and interviews in order to collect data.
- Identify and examine collection development policies and related documentation.
- Synthesize findings into a report that can be used by subsequent project activities described below.

Outcomes

- Effective outreach to practitioners tasked with selecting born-digital publications.
- White paper synthesizing findings from the collected data.
- Suggestions of characteristics that could be used to inform feature extraction and selection for training machine learning models.

Work Area 2: Building Training Datasets Work with the combined project team to construct training data sets for subsequent machine learning model building.

Tasks

- Identify PDF and DOC files from Web archives for three domains: institutional repositories, state publications, and federal technical reports.
- Code each resource as being of potential interest to a given domain or not being of interest.
- Document and release datasets under an open data license.

Outcomes

- Three training datasets will be constructed for use in training machine learning models, one for each domain: institutional repositories, state publications, and federal technical reports.
- These training datasets will be made available publicly under an open-source license.

Work Area 3: Processing Pipeline and Feature Extraction Informed from the previous tasks, design novel features for machine learning algorithms to automatically identify content-rich documents. Build a data processing pipeline to extract these document features.

Tasks

- Implement research infrastructure to extract files and metadata from Web ARChive (WARC) files.
- Develop and implement a generic framework for extracting novel topical, structural, text density, and layout features that are designed to incorporate aspects specific to content-rich documents from Web PDF and DOC files.
- Extend the feature extraction framework to incorporate findings from Work Area 1.

Outcomes

- Documented and published workflows for extracting file and metadata content from WARC files with existing tools from the Web archiving community.
- Documented workflows and open-source code for extracting novel topical and structural features from Web PDF and DOC files.
- Documented workflows and open-source code for adding domain specific feature extraction to the feature extraction pipeline.

Work Area 4: Building and Testing Machine Learning Models Using the constructed datasets and feature selection methods created during the project train and test various machine learning models to automatically identify documents of interest.

Tasks

- Implement a Bag-of-Words (BoW) based classifier trained using documents from training sets created in Work Area 2.
- Develop, evaluate, and fine-tune various topical and structural feature-based classifiers including Decision Tree, Naive Bayes, Random Forest, Artificial Neural Networks, and Support Vector Machines on the constructed dataset. Train and test classifiers using a train/test split or a cross-validation setting.
- Experiment with hybrid BoW, topical and structural feature-based classifiers.

Outcomes

- Bag-of-Words (BoW) based classifier models.
- Topical and structural feature-based classifier models.
- Hybrid BoW, topical, and structural feature-based classifier models.
- White paper describing the chosen extracted features and their effectiveness in classifying publications from the training set.

Work Area 5: Identifying Potential Documents from Previously Unseen Collections Using the highest performing machine learning models identified in Work Area 4, automatically classify extracted publications from three domains as being relevant to a collection or not, and manually evaluate the classification performance.

Tasks

- Apply the trained models from Work Area 4 to collections from three different Web archives.
    - UNT.edu Web archive, 2006-2017 longitudinal Web archive
    - State of Texas Web archive, 2009-2017 longitudinal Web archive
    - 2008, 2012, and 2016 End of Term Web archives
- Identify and score candidate publications from Web archives.
- Evaluate how the trained models will generalize to unseen documents. Manually verify a subset of identified documents to test the accuracy of extraction.

Outcomes

- Trained models that have been tested against the task of identifying publications from Web archives in a real world scenario.
- White paper with analysis of experiments including lessons learned and next steps, and dissemination of results by presentations in prestigious venues and invited talks.

2.3 Activity Details

Work Area 1: Knowledge Gathering In this work area, the research team will focus on qualitative data analysis. Specifically, the research team will conduct observation and in-depth interviews with practitioners

responsible for selecting documents from the Web for existing collections at their institutions. Currently, not enough is known about the process that collectors of born-digital Web publications use in making their selections. The project team's use of observation and interviews with practitioners will provide important information that would be challenging to identify with survey-based data collection methods. These practitioners will be identified with help from the project's Advisory Board and through existing professional relationships. We expect to conduct between 9 and 12 interviews, at least three each in the areas of institutional repositories, state publications and federal government information. Each interview will last 60 minutes and will be conducted either in-person or over a phone or video conferencing system. All interviews will be recorded and transcribed. The transcribed interview data will be analyzed using analytic induction, a mixture of deductive and inductive approaches (Epstein and Martin, 2004). We will develop a set of codes based on insights we gain from the larger research, previous studies, and the interview questions. This inductive approach is a typical approach to qualitative data analysis. For these codes, the process will be iterative and cyclical, drawing from a framework developed by Seidel (1998).

The project team will collect and analyze existing collection development policies from organizations represented in the observations and interviews as well as more broadly from collecting organizations including libraries, archives and museums from across the country. This documentation will help to inform the project team of how web-published documents are being planned for or included in collection development plans.

Before conducting observation and interviews, the project team will have all questions and planned interactions approved through the UNT Institutional Review Board (IRB).

Work Area 2: Building Training Datasets In order to accurately classify documents as being of possible interest for addition to existing collections using machine learning techniques, we need a gold-standard dataset of labeled documents from Web archives. To the best of our knowledge, there is no such publicly available dataset. Our second work area is to build such a dataset.

Building on knowledge gained from the first work area, and with input from domain experts, we will first derive rules for labeling documents as relevant, i.e., within the collecting scope or not within the collecting scope. Based on the derived rules, we will use the project team to manually label documents extracted from existing Web archives. A total of three datasets will be created, one each for the broad areas of interest to this project that include: institutional repositories, state publications, and federal publications. We will confirm the agreement between annotators assigning the labels to ensure that inter-rater reliability is sufficient to be used in training machine learning classifiers.

Work Area 3: Processing Pipeline and Feature Extraction In this project, we propose novel features that are designed to incorporate aspects specific to content-rich documents from Web PDF and DOC files. These aspects include keyword-based topicality, structural, text density and layout characteristics.

In our previous work, we successfully used structural, text density, and layout features in conjunction with machine learning classifiers to: (1) automatically identify research publications from a crawled set of Web documents to be indexed in a scholarly digital library (Caragea et al., 2014a); and (2) automatically identify the types of documents as research papers, slides, books, theses, and resumes, for appropriate indexing in digital libraries (Caragea et al., 2016). Building directly on this original work, we will extend our models along two directions: (i) we will extend our structural, text density, and layout features to capture specifics of documents relevant to institutional repositories, state library and Web archives, and End of Term Web Archives; (ii) we will provide solutions to handle the very high dimensional spaces of the Bag-of-Words representation and still capture aspects related to topicality. In our previous work, we developed highly accurate supervised and

unsupervised approaches to keyphrase extraction (Florescu and Caragea, 2017; Sterckx et al., 2016; Gollapalli and Caragea, 2014; Caragea et al., 2014b). We will leverage these lines of work on document identification and keyphrase extraction to gradually identify documents of interest to be added to a given collection type, starting with documents relevant to institutional repositories, and ending with those relevant to government information collections.

Work Area 4: Building and Testing Machine Learning Models Using our project-generated datasets, we will extensively test our proposed models with respect to at least the following criteria: (1) time and space efficiency and (2) classification performance of the trained models. Evaluation will be conducted using k-fold cross-validation on the labeled document datasets, or using a train/test split, for each of the three domains, institutional repositories, state publications, and federal technical reports. In all scenarios, we will compare our results (i.e., the predictions) against the ground-truth (i.e., human labeled) documents from our collections. In k-fold cross-validation, a dataset is split in k subsets of approximately equal size and k-1 subsets are used for training, whereas the remaining subset is used for testing. The procedure is repeated k times with each subset being considered a test set in one of the k iterations. In order to tune model hyperparameters (e.g., the number of trees in a Random Forest classifier), we will use a held-out development set, which will be sampled from the training set.

Work Area 5: Identifying Potential Documents from Previously Unseen Collections In addition to the evaluation of our classifiers on our labeled datasets, we will investigate how well our best performing classifiers developed in Work Area 4 will perform "in the wild." In particular, we will use our classifiers to identify documents of interest from very large unlabeled collections of documents from all three domains in our study, including UNT.edu Web archive; State of Texas Web archive; and 2008, 2012, and 2016 End of Term Web archive. We will then draw a random sample from the predicted documents and manually verify the accuracy of the predictions. Through this analysis and using input from librarians and domain experts, we will tune our models to meet the accuracy and error detection rate deemed acceptable for reliable collection construction.

2.4 Project Management

Project Team

This project is a collaboration between the UNT Libraries and the UNT Computer Science and Engineering Department. As such the responsibilities for leading and managing the grant will be shared between PIs Mark Phillips and Cornelia Caragea.

Mr. Phillips will serve as Principal Investigator for the project. He has extensive experience in grant-funded projects for digital libraries and Web archives as well as experience in grant-funded research projects involving ethnographic-based research methods including observation, interviewing and focus groups. His responsibilities will include: overall project supervision and budget oversight; editing and submission of required reports and grant documentation; participation in project meetings; drafting project reports; and official communication with IMLS. Phillips will be responsible for leading the qualitative research in the knowledge-gathering work stream, supervising one of the graduate research assistants, and coordinating external project communication and outreach.

Dr. Caragea will serve as Co-Principal Investigator for the project. She has an extensive background in the areas of machine learning and natural language processing: she has worked on numerous externally funded projects in a variety of roles at Pennsylvania State University and the University of North Texas. Her project responsibilities will include developing machine learning methodologies used by the project, supervising the

Computer Science and Engineering graduate research assistant, performing data analysis and evaluation, and drafting project reports and publications.

The project will hire two **research assistants**, one working primarily with Phillips and one working primarily with Caragea. It is expected that the research assistants will come from the College of Information's Information Science Department and the College of Engineering's Computer Science and Engineering Department, respectively. The research assistants will have major roles within the project including data collection, transcription and coding, development and implementation of feature extraction workflows, tuning of algorithms, and assistance in writing of white papers and research reports.

In addition to these four primary participants in the project team, the project will make use of the experience of **Lauren Ko**, head of the Software Development Team in the Digital Libraries Division. Her experience with the construction and intricacies of Web archival data formats and infrastructure will be an important component of the project.

Advisory Board

This research project will make use of an external advisory board with members from a wide range of institutions and backgrounds that will assist in guiding the project to successful completion. All members of the advisory board have deep experience in one or more aspects of the research project. The following individuals have already committed to serving on the advisory board for this project: Jefferson Bailey (Web archives), Lee Giles (digital libraries, information retrieval), Valerie Glenn (federal publications), Raymond Mooney (machine learning), Mark Myers (state publications), Maliaca Oxnam (technical reports, scholarly communication), Sarah Shreeves (institutional repositories), and Andrea Tapia (ethnographic research methods). See attached letters of commitment for a more in-depth discussion of their interest in the project. These advisory members will participate in virtual meetings spaced throughout the project timeline. The goal is to convene virtually at least two full advisory meetings per year of the grant period. These meetings will allow the project team at UNT to solicit feedback related to drafts of white papers and preliminary research findings. Advisory board members will help to identify practitioners who would be good candidates for observation or interviewing during the project.

2.5 Project Dissemination and Sustainability

This project will create and maintain a project website in the form of a blog to discuss the goals of the research, progress in the research, and to solicit suggestions and comments on working papers and other documents. In addition to this blog we will make use of a number of existing communication platforms at UNT including the UNT Libraries news site, the UNT Libraries Twitter account, and similar outlets in the Computer Science and Engineering Department. All software and technical products created by the project will be released under open-source licenses and published on the UNT Libraries' GitHub (https://github.com/unt-libraries/) code-sharing platform. The project will publish synthesis of observations and interviews, technical documentation, and all white papers in the UNT Scholarly Works Repository (https://digital.library.unt.edu/scholarlyworks/). All publications will be released under Creative Commons BY-NC-SA license. The project team will work collaboratively to submit proposals for presentations and papers at appropriate conferences and events throughout the project.

In regard to sustainability, this research project leverages existing expertise, collections, and infrastructure at the University of North Texas. The UNT Libraries has a record of not only conducting research, but widely disseminating the results and findings through conferences, articles, and white papers. It is expected that if

successful this research will lead to future research and demonstration grants related to extracting documents from Web archives.

3. National Impact

The proposed project will have national impact through investigating the feasibility of using machine learning algorithms and trained models to classify publications from Web archives automatically. Select outcomes of specific national impact are:

- A whitepaper describing collection development decisions and scoping from our identified collection curators
- Datasets informed by the interviews with collection curators that can be used for training and testing machine learning algorithms to identify and extract documents and publications from Web archives
- Feature extraction workflows and methods that are identified as informative for document classification.

Our proposed project aligns with an IMLS strategic plan goal, from *Creating a Nation of Learners:* "IMLS supports exemplary stewardship of museum and library collections and promotes the use of technology to facilitate discovery of knowledge and cultural heritage." Additionally, the project fits well into the National Digital Platform program in that the output of our research can directly assist a large number of institutions of all sizes across the country in further curating their Web archive collections. The long-term goal of this line of research is to develop workflows and methods for cultural heritage institutions to leverage their Web archive collection in new ways, such as extracting content-rich publications that would have traditionally been included in their collections but which are sometimes challenging to identify in a large Web archive.

This project is the first step in a series of research and demonstration projects aimed at improving the reuse of and access to content collected as part of Web archiving initiatives around the country. The three primary use cases highlighted in this project exemplify potential applications of this research but there also exist more focused collection building possibilities such as building collections of publications on subjects like climate change or environmental policy. It is expected that if this approach to use machine learning algorithms and models to identify publications for digital collections is successful, then there will be subsequent demonstration projects to show how this approach can be introduced into production workflows and further generalized for wider adoption and implementation. Additionally, with a potentially high number of documents being extracted from Web archives, new workflows and tools may be needed to curate and create descriptive metadata for these resources in order to provide high-quality access to our users.

A small but important feature of this project is to help cross-pollinate the library and machine learning communities through graduate research assistants in each of these areas. A goal is that these students will be better able to appreciate the work in the other areas within the project. Finally, exposing future researchers and practitioners to the technical side of Web archives and collection building will benefit the field in general.

The reuse of Web archives in building more discoverable collections is an area of significant future research potential. The ability for libraries, archives, and museums to identify and extract publications from previously collected content in Web archives is an important first step in fully realizing the potential of Web archives and digital libraries in a broader National Digital Platform.

University of North Texas

Programmatic Extraction of 'Documents' from Web Archives — Schedule of Completion

Dec 1, 2017 - Nov 30, 2019

	2017	2018				2019			
Activities & Milestones	Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Nov
Project Management & Oversight									
Develop & maintain work plan *Lead: Phillips, with Caragea*									
Create & maintain Web presence *Lead: Phillips and Caragea*									
Hire graduate research assistant Libraries *Lead: Phillips*									
Hire graduate research assistant Computer Science *Lead: Caragea*									
Software Development Guidelines and Release Process *Lead: Phillips and Caragea*									
Advisory Team Conference Calls *Lead: Phillips*									
Work Element 1: Knowledge Gathering									
Identify Participants for observation and interview *Lead: Phillips*									
Conduct observation and interviews *Lead: Phillips, with Caragea*									
Identify, Collect and Examine policies and documentation *Lead: Phillips, with Caragea*									
Synthesize findings into report *Lead: Phillips*									
Work Element 2: Building Training Dataset									
Identify PDF and DOC files from three existing datasets *Lead: Phillips*									
Code identified documents *Lead: Phillips*									
Document and release datasets *Lead: Phillips, with Caragea*									

	2017	2018				2019			
Activities & Milestones	Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Nov
Work Element 3: Processing Pipeline and Feature Extraction									
Develop local research infrastructure for WARC files *Lead: Phillips, with Caragea*									
Develop and implement generic feature extraction workflow *Lead: Caragea, with Phillips*									
Extend feature extraction to incorporate Work Area 1 findings *Lead: Caragea, with Phillips*									
Work Element 4: Building and Testing Machine Learning Models									
Implement bag-of-words (BoW) classifier *Lead: Caragea*									
Implement feature-based classifiers *Lead: Caragea*									
Implement hybrid Bow, topical and structural feature-based classifiers *Lead: Caragea*									
Work Element 5: Identifying Potential Documents									
Apply trained models from Work Area 4 to existing Web Archive collections *Lead: Caragea, with Phillips*									
Identify and score candidate publications from Web archives *Lead: Caragea, with Phillips*									
Evaluate trained models on unseen documents. *Lead: Caragea, with Phillips*									
Reports:									
Draft, review, and publish white paper with findings from all Work Areas. *Lead: Phillips and Caragea*									
Draft, review, and publish white paper outlining workflow for extracting content-rich documents from Web archiving using machine learning methods. *Lead: Caragea and Phillips*									

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products 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 can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, 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
You must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

PART I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

The products produced as outcome of our proposed effort are white papers, studies, datasets, trained models and software.

Software created during the project will be released under a BSD 3 Clause License

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The University of North Texas will assert ownership of software, documentation and research output of this project in accordance with the UNT Research Intellectual Property Policy 08.003 (https://policy.unt.edu/policy/08-003).

All output of this research project will be licensed using common open licenses including BSD 3 Clause License for software products, Creative Commons BY-NC-SA licenses for white papers, project documentation, and website, and

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

As part of this project, we will be conducting interviews and taking notes during ethnographic observations. The data collected via interactions with human subjects will be stored securely and accessed by project investigators only. Such data will be shared only after appropriate anonymization or with explicit consent from participants.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

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

New Digital Content, Resources, and Assets will not be created as output of this research project.

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

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

B. Workflow and Asset Maintenance/Preservation

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

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and 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 C.F.R. § 200.461).

C. Metadata

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

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

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets 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).

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

Part III. Projects Developing Software

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) it will serve.

While the goal of this research project is not to create production tools and services, some software will inevitably be created to assist in the research. This software will generally consist of scripts to process content from Web archives from standard WARC file formats and extract features from PDF and Word Document files. It is a goal of the project to reuse as many existing tools for this work as possible and only customize or write original software when necessary. Th

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

As noted above, while this project will produce some software that will be released the primary goal of the project is not a software development project but a research project. As such the research workflows developed for this project will make use of existing components whenever possible. Software will be developed to integrate these components into a research workflow for the research project. To our knowledge there is no software that exists for this exact work, though the project

B. Technical Information

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

Software for this project is in Java, Python, and shell scripts.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

Software developed in this research project will make use of existing tools in the areas of Web archives, natural language processing, keyphrase extraction, and machine learning. We will make use of existing software tools and libraries in these areas and combine them in different ways to develop workflows that we use for the project. In the course of the project if there are contributions to these existing libraries that can be made that are within scope and budget of the project we will

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

Software scripts designed as part of this project will be expected to run in a command-line environment and will be tested on both Linux and MacOS operating systems. The project will author software in Python, Java and Shell scripts that are available on most systems. The research workflows will draw upon existing software libraries written in C, Java, and Python that will have additional required dependencies. All expectations for the computing environment will be documented as part of the software development process.

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

A repository for the project will be created on the UNT Libraries GitHub (https://github.com/unt-libraries/) where software for the project will be kept. Initial development of scripts will be done in local development environments and then deployed from GitHub into computing infrastructure used for processing in this project. We will use the documentation and issue-tracking services provided in the GitHub repository or maintaining and updating documentation for the

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

The UNT Libraries maintains a number of publicly available and Open-Source Licensed software projects on its GitHub code sharing site (https://github.com/unt-libraries/). Projects of note include:

PREMIS Event Service - https://github.com/unt-libraries/django-premis-event-service

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? 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 and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

The University of North Texas will assert ownership of software and documentation of this project in accordance with the UNT Research Intellectual Property Policy 08.003 (https://policy.unt.edu/policy/08-003). The University of North Texas will license all software and documentation developed during this project with an open-source software license.

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

The source code for scripts developed during this research project will be made available via the UNT Libraries GitHub code sharing repository (https://github.com/unt-libraries/). Links to available software will be added to the project's website/blog so that interested parties and potential users can better find the software and documentation.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository: GitHub

URL:https://github.com/unt-libraries/

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

Data for the knowledge-gathering work stream will be collected via phone interviews, in-person interviews, and ethnographic observations, which involve note-taking, recording, and photographs. Interviews and observation will be conducted at the beginning of the project and continue through the first six months of the project. Follow-up interviews and additional recordings of conversations and note-taking will take place throughout the project as a need to document

A.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?

Data collection during the Knowledge-Gathering Work Stream involves human subjects and requires IRB approval. IRB application will be prepared and submitted when/if the project is approved for funding.

Data compiled into datasets for use in training machine learning algorithms does not require IRB approval.

A.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).

Participants can be identified in interviews, notes, and recordings. Personally identifiable information will be stored securely and only Principal Investigator and co-Principal Investigator will have access to it. Before public release of the dataset all PII will be removed (participants will be assigned coded numbers and any information that may identify them individually will be obscured in the interviews, notes, and transcripts).

A.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.

Participants will be provided with informed consent forms, which they will sign. The forms will be stored securely and separately and the relationship to the collected data will be maintained via a study ID that will be recorded in the informed consent forms and in the data files.

A.5 What methods 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).

The data will be collected via interviews and observations and will consist of text files, audio and video files, and photographs. Common word processing software and multimedia players may be used to display the data. Processed data may consist of additional spreadsheets and visualizations, which will be stored in non-proprietary formats (e.g., CSV or PNG).

A.6 What documentation (e.g., data documentation, codebooks) 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?

Codebooks will be created as part of the analysis of qualitative data (e.g., in the thematic coding procedures codes will be developed in the inductive manner, after close iterative reading of the interviews). Codes, their descriptions and other documentation that describes when and where the interviews and observations took place will be stored in text formats along with the data. The documentation will be associated with the datasets through consistent file naming and through

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

Datasets created during this project will be managed and archived in the UNT Data Repository, a collection in the UNT Digital Library operated by the UNT Libraries. All content added to the UNT Data Repository will be preserved and made available by the UNT Libraries in perpetuity.

A.8 Identify where you will deposit the dataset(s):

Name of repository: UNT Data Repository

URL: https://digital.library.unt.edu/datarepository/

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

Principal Investigators will monitor the implementation of this data management plan. The plan will be reviewed every 6 months and adjusted according to the amount and types of data generated.