Leveraging Existing Bibliographic Metadata to Improve Automatic Document Identification in Web Archives

The University of North Texas (UNT) Libraries in partnership with the University of Illinois Chicago (UIC) Computer Science Department is seeking IMLS support for an applied research grant that aligns with the agency Objective 3.1 to support collections care and management and the program's Objective 3.2 to support innovative approach to digital content management with the long-term objective of improving access to digital resources housed in web archives. This applied research project will build on findings from a previously funded IMLS research grant (LG-71-17-0202-17) that was a first effort in training machine models to help identify high-value documents and publications within web archives. This project seeks to incorporate existing bibliographic metadata related to state government document collections to better train machine learning models and allow for a reduction in human effort, as the process is still time consuming and requires highly-trained content curators. The project team includes PI Mark Phillips, Associate Dean for Digital Libraries at the UNT Libraries, and Co-PI Cornelia Caragea, Associate Professor in the UIC Computer Science Department. Two graduate students will assist in the project. The project team will partner with the Library of Michigan and the Internet Archive's Archive-It service as data sources to further test this new approach in building machine models for this task. Finally, an advisory committee of professionals from collecting institutions and machine learning researchers will provide guidance and advice for the project. We respectfully request $385,769 in support.

Project Justification

Web archives have continued to gain popularity in collecting institutions like libraries, archives, and museums. These harvested websites serve a wide range of needs, including the preservation of important publications and documents for an institution's holdings. While accessible through the web archive itself, these materials are often not identified or described at the item level, preventing users from finding resources and institutions from maintaining complete collections. As more web content is collected, identifying resources within a web archive that meet an existing collection development scope becomes more costly and daunting. For example, finding an annual report from a state agency in a web archive currently requires knowing where that report was published, and trying to find earlier years of the report requires trial and error with a web archiving interface. This access pattern is different from the curated library interfaces that allow for browsing of documents using metadata fields such as publication date, title, or publishers. Enabling libraries and archives to identify and extract documents and publications from web archives to include in existing library collections is still an activity that is desired by groups working in this space (Fox et al., 2020)

In 2017, the UNT Libraries and the UIC Computer Science Department received IMLS support under the National Digital Platform category for a research project to evaluate the use of machine learning algorithms to identify and extract publications contained in existing web archives as a way of surfacing these documents. This research was remarkably successful, resulting in: (1) the development of three gold-standard, manually-labeled datasets from three different web archive domains, an institutional repository from a web archive of a university domain, state publications from a web archive of a state government, and technical reports from a large federal agency; and (2) the design of supervised machine learning models that accurately classify PDF documents from web archives against the scope for a given collection. We disseminated this research in prestigious venues in digital libraries and natural language processing and won The Best Student Paper Award on one of our papers published in the Joint Conference on Digital Libraries (JCDL) in 2020.

Despite this success, we identified three major challenges that we aim to address during this project. First, one of the biggest challenges is that training accurate supervised machine learning models requires human-annotated data which are expensive to obtain and often impractical as human annotators need a certain amount of background knowledge and domain expertise to perform the annotation task. An example of this human-annotation is the labeling of documents extracted from a web archive as either “in-scope” or “not-in-scope” based on an existing collection area such as state publications. Second, although unlabeled data (i.e., large collections of unlabeled documents) can be obtained almost for free using methods like web archiving, it is unclear how to successfully use these unlabeled data to build accurate machine learning models, while reducing human effort. Third, it is unclear how the built models generalize on data “in the wild” (e.g., data from a different state) and how robust these models are under distribution or vocabulary shifts without the need to repeat the entire annotation-training pipeline for a new state (e.g., on data from one state to another under vocabulary distribution shifts, or from one collection type/scope to another), when no labeled datasets are available in the new state.

We are now seeking IMLS support for a research project to address these challenges through the following tasks: (1) explore the use of available bibliographic metadata (representing decades of work by librarians within the defined collection scopes) together with unlabeled data to create large machine-learning training sets while reducing human effort; (2) focus on the design of machine learning algorithms for incorporating information from unlabeled data using positive-unlabeled learning techniques; and (3) focus on the design and development of unsupervised domain adaptation techniques, i.e., training on one domain (the source domain) and testing on another (the target domain), to address model generalization under distribution or vocabulary shifts. As a by-product of this research, we will generate test datasets from curators at the Library of Michigan and their web-archived content via the Internet Archives’ Archive-It service that will serve as evaluation test sets for in-domain and out-of-domain scenarios.

Project Work Plan

Our research project tests the assumption that we will be able to successfully incorporate existing bibliographic metadata and catalog records to aid automated classification algorithms in accurately extracting documents and publications that meet collection development policies for a variety of organizations. We will focus on state government documents held in web archives at the UNT Libraries (for Texas) and at the Library of Michigan (for Michigan). We will use existing bibliographic metadata at these institutions, which describe and implicitly encode collection decisions from decades of manual curation efforts.

During the project, the team will investigate weak supervision by mapping bibliographic metadata onto existing datasets, annotated during our previous/completed research, to understand which metadata fields are highly correlated with the positive vs. negative class and exploit these correlations to automatically create large model-training sets. Additionally, the team will treat the task as positive-unlabeled learning, in which an incomplete set of positive examples is available as well as a set of unlabeled examples (some positive and others negative). We will adapt the learning algorithm so that training examples have individual weights, i.e., positive examples are given unit weight and unlabeled examples are duplicated where one copy of each unlabeled example is made positive with some weight w and the other copy is made negative with weight $1-w$. By reweighting the importance of these training examples, we aim to model the uncertainty in the negative examples. Finally, we will focus on the design of unsupervised domain adaptation models to understand how well models trained in one domain (Texas State Publications) can be adapted to another domain (Michigan State Documents). Additional details of these activities are provided below.

Activity Overview

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.

Project Goal: The overarching goal of this project is to investigate the potential of using existing bibliographic metadata related to state government document collections to better train machine learning models that can assist librarians and information professionals in identifying and classifying high-value publications from large web archives. This integration will hopefully result in a reduction in human effort, as the process of creating large labeled datasets that are required for training machine learning models is still time consuming and requires highly-trained content curators.

We envision that 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. *How can large amounts of training data be generated for supervised approaches with less intensive human effort, which is often impractical?*
2. *How can we successfully incorporate information from unlabeled data to build robust classifiers for identifying documents in-scope of a collection?*
3. *How will our models generalize to data "in the wild" (i.e., data from a different state) and how robust are the models under distribution or vocabulary shifts (e.g., on data from one state to another under vocabulary distribution shifts, or from one collection type/scope to another), when no human-annotated datasets are available in the new / target domain?*

Research Methodologies: The project will make use of research methodologies and practices from the domain of library and information science such as metadata normalization and normalization with machine learning techniques from computer science to answer the above research questions. The multidisciplinary 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 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. Additionally, arrangements have been made with the Library of Michigan to use their existing bibliographic metadata for their state publications as well as existing web archives that have been collected and are housed with the Archive-It service at the Internet Archive. Letters of support from these institutions in addition to representatives being present on the advisory board help assure access to the necessary data for the successful completion of this project.

Assumptions: Because this project builds on previous research conducted successfully by the project team, there is a baseline of knowledge on the feasibility of extracting high quality documents and publications from web archives at scale. Assumptions in this work arise from the expectation that in order to successfully extract publications from different domains, for example different state publications, we will need to tailor the models to that specific domain in some way. Our base assumption is that we will be able to successfully incorporate existing bibliographic metadata and catalog records to aid automated classification algorithms in accurately extracting documents and publications that meet collection development policies for a variety of organizations.

The following tables help to summarize the five different work areas of this research project and their associated tasks and outcomes. Additional details are presented under each work area below. Further information about the sequencing of work areas and the overall progression of the project are presented in the included Schedule of Completion.

Work Area 1: Data acquisition and pre-processing

Tasks and Outcomes for Work Area 1
Tasks ● Acquire UNT Libraries' Texas State Publications metadata from library catalog and digital collections. ● Extract PDF data from texas.gov crawls at UNT. ● Acquire Library of Michigan's State Publications metadata from library catalog and digital collections. ● Acquire Library of Michigan's web archives from Archive-It. ● Extract PDF data from Michigan web archive.
Outcomes ● Dublin Core metadata dataset for Texas State Publications ● Extracted PDFs and associated request/response information for Texas web archive ● Dublin Core metadata dataset for Michigan State Publications ● Extracted PDFs and associated request/response information for Michigan web archive

This research proposal requires several new datasets to be built from raw data that exists in multiple locations and in multiple formats. First bibliographic metadata from library catalogs and digital collections will be aggregated. These collections of metadata will be gathered from both MARC-based catalog records that describe previously collected and described resources that fit the collection scopes of an institution. In this project, two datasets of bibliographic metadata will be gathered. The first set will be from the University of North Texas Libraries' catalog. At this time there are over 13,700 bibliographic records from the Texas State Documents Collection in the UNT Libraries' Discover system https://discover.library.unt.edu/ that will be exported in the MARC format to use in training data. The 17,000+ metadata records from the Texas State Publications collection in The Portal to Texas History https://texashistory.unt.edu/explore/collections/TXPUB/ will be downloaded using the existing OAI-PMH repository endpoint. Both of these sets of bibliographic metadata will be converted into the Dublin Core format for a consistent representation of the underlying data and organized for processing in other Work Areas of this project. An extracted collection of PDF files from the UNT Libraries Web Archive from previously conducted texas.gov focused web crawls will be pulled from WARC files. In addition to the PDF file itself, metadata about the crawl will be extracted, including request and response headers that were generated during the archiving process. The combination of this PDF data and the normalized Dublin Core metadata from the UNT Libraries Catalog and the Texas State Publications collection in The Portal to Texas History will complement the existing datasets described above from previous funded work in this area. These combined datasets will be used in Work Area 2 and Work Area 3.

To assist with Work Area 4 and Work Area 5, a similar dataset will be constructed from holdings of the Library of Michigan. This will mirror the data created for Texas and will include MARC-based bibliographic metadata from the Library of Michigan's catalog https://magic.msu.edu/search~S37 and harvested metadata from the Library of Michigan's Digital Collections https://lmdigital.libraryofmichigan.org/ related to Michigan State Publications. These different bibliographic metadata collections will be normalized into the Dublin Core metadata format in a similar way as the Texas metadata datasets. The web archives collected by the Library of Michigan using the Internet Archives' Archive-It tool, called the Michigan Government Web Collection https://archive-it.org/collections/418, will be acquired in collaboration with Archive-It and with the permission of the Library of Michigan. This will be done using the Web Archiving Systems API (WASAPI) data transfer APIs developed in a previous IMLS-funded National Leadership Grant (LG-71-15-0174).

WASAPI allows for the systematic harvesting of web archives, including those collected by the Archive-It service. After harvesting these WARC records, the PDF content from the collection will be extracted and organized in a format identical to the Texas dataset mentioned above.

Work Area 2: Using weak supervision to reduce the human effort for data annotation

Tasks and Outcomes for Work Area 2
Tasks ● Map the available bibliographic metadata onto unlabeled web archiving collections to create weakly supervised large machine-learning training sets while reducing human effort. ● Identify what metadata fields (e.g., title, creator, subjects) are most useful for obtaining weakly-annotated training sets using our existing human annotated datasets from our prior work. ● Train baseline classifiers including Decision Tree, Naive Bayes, Random Forest, Support Vector Machines, Deep Neural Networks, and Pre-trained Language Models such as BERT on these weakly-annotated datasets, where the mapped documents are the positive examples and the remaining documents are the negative examples.
Outcomes ● Weakly annotated training datasets for use in training machine learning models from the Texas State Publications domain and associated bibliographic metadata, made available publicly under an open-source license ● Metadata fields most useful for weak annotation for each of the three domains ● White paper describing the best metadata fields and their effectiveness in creating training datasets with reduced human effort and the best performing classifiers trained to identify documents of interest to a collection using the created training sets

In our previous work (Patel, Caragea, & Phillips, 2020; Patel, Caragea, Phillips, & Fox, 2020), we created three gold-standard datasets of labeled documents from web archives from three different web archive domains https://www.cs.uic.edu/~cornelia/datasets/web_archive_data: an institutional repository from a web archive of a university domain, state publications from a web archive of a state government, and technical reports from a large federal agency. While these datasets helped facilitate the successful use of machine learning to automatically classify documents as being of interest for addition to existing web archive collections, it was evident that significant amounts of human effort and time were required to annotate them. In addition, from our prior research, we learned that one needs a certain amount of background knowledge, domain expertise, and training to annotate such datasets. However, with the diverse and numerous web archive collections (e.g., from different states such as Texas and Michigan, and from different domains such as a web archive of a university domain or technical publications from a web archive of a federal governmental agency), using human annotators to annotate training datasets for each of these collections is not scalable nor sustainable.

In this work area, we will explore the use of available bibliographic metadata (representing decades of work by librarians within the defined collection scopes) as weak supervision together with unlabeled data to create large machine-learning training sets while reducing human effort. Specifically, the available bibliographic metadata consists of fields such as publisher, agents [publisher + creator + contributor], title, description, or subjects/keywords about the documents of interest for a web archive collection (i.e., the positive class). We will map these metadata records against large unlabeled document sets from a web archive collection to obtain weakly annotated training data. Let us consider as an example the web archive of Texas websites at the UNT Libraries (texas.gov). With the metadata information available from librarians for the documents of interest to the collection, e.g., title, keywords, or creator (or a combination of these), we will search the large unannotated document collection crawled from the texas.gov domain. If a document from this collection matches (or approximately matches, up to a certain threshold) the searched metadata information, we will label the respective document as positive; otherwise, we label the document as negative). This task, known as record linkage, is very challenging since even a small difference in the compared strings (the string from the metadata record with the corresponding string from the documents in the collection) would result in a mismatch. In our previous work, inspired by Bhattacharya and Getoor (2004), we successfully performed record linkage by approximate matching in order to build a scholarly dataset of clean research papers and automatically removed noise in a digital library collection (Caragea et al., 2014).

To understand the efficacy of our weak supervision approach in this project, we will evaluate this on our three human annotated datasets from our previous work for which we know the correct labels to determine the level of potential label noise introduced through weak supervision. In addition, we will train machine learning and deep learning classifiers on these weakly annotated training sets and compare their performance with that of counterpart classifiers trained on our human-annotated collections. These classifiers will represent our baseline classifiers and will include Decision Tree, Naive Bayes, Random Forest, Support Vector Machines, Deep Neural Networks, and Pre-trained Language Models such as BERT (Bidirectional Encoder Representations from Transformers).

Work Area 3: Positive-Unlabeled Learning

Tasks and Outcomes for Work Area 3

Tasks

- Formulate the task of identifying documents of interest to a collection as positive-unlabeled learning, where the positive set consists of the mapped documents from bibliographic metadata, with the remaining documents being considered as unlabeled.
- Develop and implement the two-stage approach of positive-unlabeled learning.
- Compare the classifiers trained using the two-stage positive-unlabeled learning approach with those trained on Work Area 2 (where we assume all unlabeled examples are negative, not unlabeled).

Outcomes

- A new approach for learning traditional binary classifiers given a nontraditional training set (of positive and unlabeled data) that is guided by a model's training dynamics
- Performance differences between the positive-unlabeled learning classifiers and the traditional classifiers trained in Work Area 2
- White paper presenting the positive-unlabeled learning framework for the classification of web archiving documents in-scope of a collection

The training data for a machine learning algorithm that learns a supervised binary classifier normally consists of two sets of examples, one set of positive examples (i.e., the documents of interest to a collection in our scenario) and the other set of negative examples (all other documents) as we constructed in Work Area 2. However, the available bibliographic metadata consists of a random sample of positive examples (incomplete by no means) with no cataloging of documents of no interest to a collection (i.e., the negative set). Thus, in our domain, the concept of negative examples is not necessarily natural. Within our weak supervision approach from Work Area 2, we assume that the documents identified based on the available metadata are positive examples, while the remaining documents from the unlabeled collection of documents are negative examples. In fact, these remaining documents (those not mapped with bibliographic metadata records) are unlabeled examples, not negative examples, because some of these documents are of interest to a collection despite that they do not appear in the bibliographic metadata files. Thus, the assumption that all remaining documents are negative introduces some errors in annotation (i.e., for the documents that are in fact positive but are not labeled as positive because they did not appear in the available metadata records or because the weak supervision failed to find a match between the metadata and any document in the unlabeled document set). It is known that in fully supervised learning, noisy or erroneous labels are problematic and can negatively impact a model's generalization performance, especially for deep neural networks, which can attain zero training error on any dataset (Zhang et al., 2021). In this work area, we aim to discover these documents and train more robust models using a novel positive-unlabeled learning approach. Precisely, we will treat the available training data as an incomplete set of positive examples (those identified using metadata information) and a set of unlabeled examples, some of which are positive and some of which are negative. Our goal will be to learn a traditional binary classifier given a nontraditional training set (of positive and unlabeled data).

To achieve this, we propose a two-stage approach: in the first stage, we train a classifier as in Work Area 2 and use it to make predictions only on the unlabeled set to reweigh the importance of these training examples. By doing so, we aim to model the uncertainty in the negative examples (in that some of them are indeed positive). The reweighted training collection together with the positive labeled set (by bibliographic metadata) are then used to train a second classifier to predict the binary labels of the documents (as being of interest to a collection or not), while at the same time automatically identifying highly ambiguous or potentially mislabeled examples that hurt model performance and generalization. Specifically, to automatically identify and remove ambiguous or potentially mislabeled examples from the training dataset, we will monitor and exploit differences in the training dynamics of clean and mislabeled examples (Pleiss et al., 2020). Intuitively, if for an arbitrary example, there is a constant tension or disagreement during training between the model predictions and the gold label of the example, then the example will more likely be mislabeled or highly ambiguous, whereas if the model predictions consistently agree with the gold label, then the example will likely be correctly labeled. Removing the harmful mislabeled examples will lead to reduced memorization and improved generalization.

Work Area 4: Validation on the Michigan Data

Tasks and Outcomes for Work Area 4
Tasks ● Leverage weak supervision and positive unlabeled learning approaches on the data from the State of Michigan to determine whether our approaches generalize or not on data from another state and identify the most useful metadata fields. ● Train classifiers on these weakly supervised training sets and evaluate them on the test sets developed as part of the Work Area 1.
Outcomes ● Weakly supervised training sets for training accurate classifiers using the positive-unlabeled learning framework ● Validation of what metadata fields are useful for weak supervision and consistency among the two states under study

In this work area, our goal is to validate the weak supervision and positive unlabeled learning approaches on the data from the State of Michigan to determine whether our approaches generalize or not on data from another state. In particular, we will make use of the bibliographic metadata from the Library of Michigan's catalog and harvested metadata from the Library of Michigan's Digital Collections related to Michigan State Publications to weakly annotate large unlabeled web archiving collections of documents as positive-unlabeled data (by mapping the bibliographic metadata onto the unlabeled data to obtain the positive examples with the remaining examples/documents being unlabeled). As already mentioned, we will acquire the web archives collected by the Library of Michigan using the Internet Archives' Archive-It tool, called the Michigan Government Web Collection https://archive-it.org/collections/418, in collaboration with Archive-It and with permission of the Library of Michigan using the Web Archiving Systems API (WASAPI) data transfer APIs. These data (the bibliographic metadata and the unlabeled web archiving collections of documents) will be used to validate our weak supervision and positive-unlabeled learning approaches on data from the State of Michigan.

During training, we will directly utilize the mapped positive-unlabeled data, which may contain some noise due to potential errors in mapping. However, in our validation scenario on data from Michigan, to have a realistic evaluation of the models' performance, we will make use of the manually (human) annotated data from Work Area 1. These clean data are used in two splits, one for development and another for testing. We evaluate our model and the baselines using standard measures for document classification such as accuracy, F1-score, precision, and recall.

Work Area 5: Model generalization for in-domain and out-of-domain data

Tasks and Outcomes for Work Area 5
Tasks ● Conduct an in-domain evaluation using a train/test split for the state publication domain and for each state, Texas and Michigan. ● Conduct an out-of-domain evaluation to derive an improved understanding of how the models will generalize to unseen documents under distribution or vocabulary shifts. ● Compare the trained models in terms of time and space efficiency as well as classification performance in both in-domain and out-of-domain settings.
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

In this evaluation, we will extensively test our models with respect to at least the following criteria: (1) time and space efficiency; (2) classification performance of the trained models when training and testing the models on data drawn from the same distribution (i.e., the in-domain setting) which means training and testing the models on data from the same domain such as a collection of state documents from the State of Michigan; and (3) classification performance of the trained models when we train and test the models on data drawn from different distributions (i.e., the out-of-domain setting) using unsupervised domain adaptation techniques, e.g., training on data from Texas (the source domain) and testing on data from Michigan (the target domain) to understand the effects of vocabulary shifts (differences in vocabularies) that may occur from one state to another or from one collection type to another, such as using an institutional repository as the source domain (the domain where the classifiers are trained) and using state publications as the target domain (the domain where the classifiers will be used to make predictions on unseen data). We will perform this unsupervised domain adaptation evaluation on data within each state and across the two states. This out-of-domain evaluation will help us understand how the models will generalize under distribution or vocabulary shifts.

For the in-domain setting, the evaluation will be conducted using a train/test split on the extracted web publications for each of the two states, Texas and Michigan. In all scenarios, we will compare our results (i.e., the predictions) against the ground-truth (i.e., human labeled) documents from our collections. For the out-of-domain evaluation of the unsupervised domain adaptation techniques, we will use the pre-trained language model BERT as the underlying classifier. Following Han and Eisenstein (2019), we will focus on using strategic pre-training techniques that will enable effective knowledge transfer between disparate domains. Specifically, the systems for comparison will be: (1) classifiers trained in the source domain and evaluated in the target domain for the same collection type, e.g., training on a repository of state documents from the State of Texas and testing on a repository of state documents from the State of Michigan; (2) classifiers trained in the source domain and evaluated in the target domain for different collection types, e.g., training on a repository of state documents from the State of Texas and testing on an institutional repository from the State of Michigan. As before, in all scenarios, we will compare our results (i.e., the predictions) against the ground-truth (i.e., human labeled) documents from our test collections.

Project Team

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

Mark Phillips, Ph.D. 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. 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 coordinating the bibliographic metadata dataset building and the acquisition of web archiving data used in the project. He will supervise one of the graduate research assistants, and coordinate external project communication and outreach.

Cornelia Caragea, Ph.D. will serve as Co-Principal Investigator for the project. She has an extensive background in the areas of machine learning, deep learning, and natural language processing (NLP): she has worked on numerous externally funded projects in a variety of roles (e.g., PI, Co-PI) at the University of Illinois Chicago and the University of North Texas. Her project responsibilities will include developing machine learning, deep learning, and NLP methodologies used by the project, supervising the Computer Science graduate research assistant, performing data analysis and evaluation, and drafting project reports and publications.

The project will hire two **graduate research assistants**, one working primarily with Phillips and one working primarily with Caragea. It is expected that the research assistants will come from the UNT College of Information's Information Science Department and the UIC Computer Science Department, respectively. The research assistants will have major roles within the project including data collection, metadata mapping and normalization, development and implementation of the weak supervision and positive-unlabeled learning workflows, tuning of algorithms, and assistance in writing of white papers and research reports.

Advisory Board

This research project will make use of an external advisory board with members from a wide range of institutions and backgrounds including both researchers and practitioners who 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 committed to serving on the advisory board for this project: Jefferson Bailey (web archives), Bernadette Bartlett (state publications), Oksana Zavalina (cataloging and metadata), Martin Klein (web archives, machine learning), Mark Myers (state publications), Tracy Seneca (web archiving, digital collections), and Raymond Mooney (machine learning). 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 quarterly 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.

Dissemination

We will systematically share four products of our research. *First*, we will share all datasets, algorithms and tools resulting from this project through GitHub https://github.com/unt-libraries and the project webpage, which will be hosted at the UNT Libraries. In particular, we will make our tools and scripts available for other researchers and practitioners to trial on their own data, and help reveal gaps in the model that do not manifest on the datasets available to us. *Second*, we will evaluate our models and tools developed during the project in the UNT Digital Library and The Portal to Texas History, which will allow users to provide feedback on the correctness of the classified documents, which will help us refine our models, in a feedback loop. *Third*, we will share our findings through publications in academic journals and presented at top conferences on library science, information retrieval, artificial intelligence, and natural language processing. Some of the venues that we plan to target include the Association for the Advancement of Artificial Intelligence, The Web Conference, Empirical Methods in Natural Language Processing, Transactions on the Web, Joint Conference on Digital Libraries, and Journal of the Association for Information Science and Technology. *Fourth,* in addition to formal journal and conference proceedings, we will work to share our project work and lessons learned with practitioners broadly in the field of library and information science. This will include submitting proposals for presentations and poster sessions at annual meetings such as the Depository Library Council, Best Practices Exchange, Texas Conference on Digital Libraries, Coalition for Networked Information, Society of American Archivists' Web Archiving Section, Web Archiving and Digital Libraries (WADL) Workshop, International Internet Preservation Consortium (IIPC) Web Archiving Conference, General Assembly, and IIPC-organized webinars. We will also work to highlight the project's activities on Twitter as part of the community hashtag of #WebArchivingWednesday https://twitter.com/hashtag/WebArchiveWednesday whenever possible to broadly increase knowledge of and about this project. All published output including datasets, articles, presentations, white papers and documentation will be available in the UNT Scholarly Works Repository and UNT Data Repository.

Project Results

Our primary goal is to test solutions outlined here in response to the challenges identified in our previous research. We hope that leveraging bibliographic metadata created over the past few decades will reduce current human effort while improving precision and recall of automated extraction of "in-scope" resources from web archives to better meet collection stewardship and user access needs.

Deliverables from this project will include: (1) approaches that incorporate existing knowledge from decades of effort encoded as metadata records in library catalogs and digital collections to create large training sets using distant supervision and positive-unlabeled data which will serve as benchmarks for training and testing *robust* machine learning algorithms to identify and extract materials from web archives; (2) accurate unsupervised domain adaptation models and workflows to evaluate model generalization under distributional / vocabulary shifts; (3) annotated test collections from a different state (Michigan) to evaluate the model generalization and transferability from one domain to another; (4) a series of white papers describing findings from this project including appropriateness of metadata fields and their effectiveness, positive-unlabeled learning framework for classification, and an analysis of experiments including lessons learned during the research.

This project will increase the understanding of how libraries and archives can make use of our existing metadata and catalog records to enable greater reuse of documents and publications captured in the web archiving process. The ability to incorporate these publications into existing platforms and discovery systems will lead to an increase in access for our users. This project will help increase our ability to identify these publications and documents in web archives.

Aug 1, 2022 - July 31, 2024

	2022		2023				2024		
Activities & Milestones	Aug-Sep	Oct-Dec	Jan - Mar	Apr - Jun	Jul - Sep	Oct - Dec	Jan - Mar	Apr - Jun	Jul
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*		🟩							
Quarterly Advisory Team Conference Calls *Lead: Phillips*		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Work Area 1: Data acquisition and pre-processing									
Acquire, normalize, and repackage UNTL Catalog and Digital Collections Bibliographic Metadata *Lead: Phillips*	🟩	🟩							
Acquire, extract and package texas.gov WARC dataset *Lead: Phillips*		🟩	🟩						
Acquire, normalize, and repackage Michigan Catalog and Digital Collections Bibliographic Metadata *Lead: Phillips*		🟩	🟩						
Acquire, extract and package Michigan WARC dataset *Lead: Phillips*				🟩	🟩				
Work Area 2: Using weak supervision to reduce the human effort for data annotation.									
Map bibliographic metadata onto unlabeled web archiving collections from Texas. *Lead: Caragea with Phillips*	🟩	🟩							
Identify and characterize metadata fields useful for data annotation with weak supervision. *Lead: Caragea with Phillips*		🟩							
Train classifiers on the weakly annotated data and compare with those trained on human annotated data. *Lead: Caragea with Phillips*		🟩	🟩						

Work Area 3: Positive-Unlabeled Learning									
Formulate the positive unlabeled learning for web archiving collections *Lead: Caragea, with Phillips*	🟩	🟩							
Develop and implement the two-stage approach of positive-unlabeled learning *Lead: Caragea, with Phillips*			🟩						
Evaluate and compare classifiers on the Texas collections *Lead: Caragea, with Phillips*				🟩					
Work Area 4: Validation on the Michigan Data.									
Explore weak supervision on data from Michigan to identify the most useful metadata fields and how well they generalize across states. *Lead: Caragea and Phillips*					🟩				
Explore positive-unlabeled learning on data from Michigan to assess model generalization *Lead: Caragea and Phillips*						🟩	🟩		
Evaluate classifiers in the unsupervised domain adaptation setting on the datasets developed as part of Work Area 1. *Lead: Caragea and Phillips*							🟩	🟩	
Work Element 5: Model generalization on in-domain and out-of-domain data.									
Train models on training sets developed for Texas and evaluate them on test sets developed for Texas (in-domain evaluation). *Lead: Caragea, with Phillips*		🟩	🟩				🟩	🟩	🟩
Train models on training sets developed for Texas and evaluate them on test sets developed for Michigan (out-of-domain evaluation). *Lead: Caragea, with Phillips*							🟩	🟩	🟩
Evaluate and compare trained models with strong baselines in terms of time and space complexity across states and across collection types. *Lead: Phillips, with Caragea*							🟩	🟩	🟩
Reports and Datasets:									
Draft, review, and publish white paper outlining workflow for leveraging bibliographic metadata to enhance models for document extraction. *Lead: Caragea and Phillips*								🟩	🟩
Package and deposit all datasets created during the grant period if not already deposited in the UNT Digital Library. *Lead: Phillips and Caragea*								🟩	🟩
Draft, review, and submit project reports. *Lead: Phillips and Caragea*					🟩				

University Libraries - University of North Texas

Organizational Profile

The University of North Texas (UNT) is a four-year public Doctoral University with a Carnegie Classification of Highest Research Activity (R1). The institution was founded in 1890 as a normal and teacher-training institute. Its name changed from North Texas State University in 1988. UNT is located in Denton, a city of approximately 121,123, and in the Dallas/Fort Worth/Arlington area of over 6.8 million. The university is accredited by the Commission on Colleges of the Southern Association of Colleges and Schools to award baccalaureate, master's, and doctoral degrees. UNT is the 5th largest university in Texas and among the 30 largest in the United States. UNT has a combined enrollment of 42,372 students and 2,148 academic staff.

The University of North Texas (UNT) Libraries are recognized both as national leaders in digitization as well as for research in the areas of digital preservation, Web archiving, and information seeking behavior. The UNT Libraries began Web archiving in 1997, when they created the CyberCemetery to capture and provide permanent public access to the Web sites and publications of defunct U.S. government agencies and commissions. In recognition of their role, the UNT Libraries exist as one of only ten Affiliated Archives of the National Archives and Records Administration (NARA).

As a member of the International Internet Preservation Consortium (IIPC), UNT has engaged in institutional, national, and international Web archiving initiatives, including the End of Term 2008, 2012, 2016, and 2020 initiatives in the U.S. Since 2010, UNT has served as a representative on the Steering Committee of the IIPC and has been involved in projects and software development activities.

The Libraries successfully managed three research projects funded under the National Leadership Program of the Institute of Museum & Library Services (IMLS): "IOGene: Interface Optimization for Genealogists" (LG-06-07-0040-07), “Classification of the End-of-Term Archive: Extending Collection Development Practices to Web Archives” (LG-06-09-0174-09), and “Programmatic Extraction of ‘Documents' from Web Archives” (LG-71-17-0202-17). The first project involved genealogists in redesigning the user interface to The Portal to Texas History, and the second project involved government information professionals in the classification of government websites in the 2008/2009 End of Term Web Archive, and the third involved developing machine models to classify documents from web archives as being in-scope or out-of-scope of existing collections of publications.

In support of their mission to support the “the long-term collection, production, maintenance, delivery, and preservation of a wide range of high-quality digital resources and services for the UNT Community and users throughout the world.” the UNT Libraries have received funding from national programs including NDIIPP, NEH, NHPRC, and IMLS. Funding from state agencies such as the Texas State Library and Archives Commission through their TexTreasures grant program. The Libraries’ own commitment to develop digital collections as core services and outreach functions is illustrated by its Digital Library Division, which includes 19 full-time staff members, 35 part-time graduate and undergraduate student employees, and the support services of the Libraries’ IT staff of 14 full-time employees.

Digital Products Plan

This applied research project will generate a number of outputs that are fully described in the Data Management Plan included in this proposal packet.

Type

This research project will produce several datasets containing normalized bibliographic metadata, extracted publications from existing web archives, and a combination of human-generated and machine-generated labeled data used in the research project.

Various software tools and algorithms will be written in order to carry out this research project.

Finally, publications, presentations and other white papers will be drafted and published during this project.

Availability

All output of this project will be made as broadly available as possible through a combination of avenues including the depositing of all datasets, whitepapers, publications, and presentations in the UNT Scholarly Works Repository (https://digital.library.unt.edu/explore/collections/UNTSW/) that is part of the UNT Digital Library (https://digital.library.unt.edu/). These items will remain a permanent part of the UNT Libraries' Digital Collections. They will be described using a metadata scheme called UNTL (https://library.unt.edu/digital-projects-unit/metadata/) that is a locally qualified Dublin Core based metadata format in use at the University of North Texas. The UNT Digital Library is heavily indexed by search engines and other scholarly aggregators and the associated metadata will improve the ability for users to discover these resources. The publications and datasets will also be listed on the project website that will be created and hosted by the UNT Libraries. This project website will provide information about the project as well as links to datasets, software tools, and publications created during the research.

We will use a Github repository to share software, scripts, and algorithms created during this research project. The Github repository will allow for interactions between the project team and others interested in using the software, scripts, and algorithms in their own work. The UNT Libraries has experience providing access to software tools and applications via Github (https://github.com/unt-libraries/)

Access

All publications and presentations will be accessible via the UNT Scholarly Works Repository. Datasets will be deposited in the UNT Data Repository (https://digital.library.unt.edu/explore/collections/UNTDRD/) that is a collection in the UNT Digital Library. Finally all software tools will be accessible via the project's Github page that will list repositories created during the research project for software, scripts, and algorithms.

Rights for all of these outputs will be assigned a broad reuse license based on the type of resource. For software, scripts, and algorithms a GNU open source General Public License (GPL). Publications and presentations, whenever possible, will be made available via a Creative Commons License such as CC-BY. Datasets will be made available under a CC0 Public Domain Dedication license.

Sustainability

By including the datasets, publications, and presentations in the UNT Libraries Digital Collections, they become part of the permanent holdings of the UNT Libraries. The long-term access to these resources is expected to be indefinite but the project team will commit to providing access to all products and data from this project for at least five years after the completion of the grant period. Software, scripts, and algorithms will be available on the Github platform as long as that is a viable and free tool for accessing software and code. In the event that it is no longer a usable platform, the team will move the repositories to another platform or archive the final version in the UNT Digital Library. Due to the research nature of this project, the software, scripts, and algorithms shared via the Github platform will be supported at least for two years after the project ends but it is not expected that they would be maintained through changes in language versions or the need to migrate to new tools.

Data Management Plan

“Leveraging Existing Bibliographic Metadata to Improve Automatic Document Identification in Web Archives” Project duration: August 1, 2022 - July 31, 2024.

This project will create several datasets that will be useful for training machine learning models to classify publications and documents extracted from web archives. These datasets will be shared widely and can be used broadly by researchers to build and evaluate new algorithms and systems. In addition to sharing these datasets, the project team will share our findings through publications in academic journals and presentations at top conferences. The Pis of this project commit to making available to the research community: databases generated during this project period, the software and tools produced, and the publications and presentations. They further commit to preserve the data online for at least five years beyond the end of the grant. All datasets and publications will be publicized on a webpage created for the project by the project team.

Datasets

Data: During the course of the project, we will generate several datasets that will be useful for other researchers and in the validation of the research outputs of this work. These datasets include normalized bibliographic datasets containing metadata records from different state publications collections. These datasets of bibliographic metadata will be derived from two sources, first traditional MARC-based catalogs that describe physical and digital state publications held by collecting institutions for two state documents collections. Second, metadata harvested from digital collections of state documents collections. These datasets will be presented in a normalized format that provides common access to the two data sources. We expect that there will be datasets created for at least two separate states during the course of this project.

In addition to bibliographic metadata datasets mentioned above, several other datasets will be generated from web archive collections of state government documents. These datasets will include links and references to extracted document files such as PDFs or DOC files that will be used as input to the machine models being researched in this grant project. These documents will be used for classification. For portions of these datasets we will also include hand and machine labeled data designating if the publications would be included or not included in a state publications collection. We expect that there will be several different datasets created in this series related to the different states we are working with on the project and will align with the normalized bibliographic metadata mentioned above.

All the above data will be made available annually in an easily-utilized format (e.g., XML, TSV, or WARC files) through the project’s website, which will be hosted on a server at UNT Libraries. These datasets will be available throughout the project and finalized versions of the datasets will be deposited with the UNT Data Repository (https://digital.library.unt.edu/explore/collections/UNTDRD/), a collection in the UNT Digital Library (https://digital.library.unt.edu). The availability of data online will make it possible for researchers working on this topic to perform fair comparisons between their algorithms and others that are developed.

Storage and Durability: We will store data on servers located at the UNT Libraries. UNIX/LINUX operating systems will be deployed on these servers for both high efficiency and high migratability. Finalized data that is deposited with the UNT Digital Library will be stored using redundant, distributed data stores as are defined in the UNT Libraries Trusted Digital Repository Self-Audit (https://library.unt.edu/digital-libraries/trusted-digital-repository/)

Strategy to support data sustainability and access: Sustainability is critical to the long-term success of the project. Because of our use of all open-source software and open standard formats, we will not incur licensing costs after the funded period of the project. Because the datasets will be deposited in the UNT Digital Libraries’ Data Repository, long-term access will be ensured as it becomes a holding of the UNT Libraries permanent collection.

Data Sharing: The PIs commit to share widely all data resulted from this project. They have been successful in making previous research available through datasets, presentations, and published literature. Furthermore, the inclusion of the finalized datasets in the UNT Digital Library will allow for standard metadata to be created and shared about the datasets that will increase the discoverability through popular data aggregators and search engines.

Algorithms and Software Tools

As part of this project, we will develop algorithms and software tools to test machine models used in classifying extracted publications from web archives. The tools will be implemented on top of existing open-source machine learning packages such as scikit-learn, PyTorch, huggingface. All of the software tools developed in this project, including the source code, will be made freely available to the research community under an GNU open source General Public License (GPL) through GitHub (github.com). In addition, software tools and documentation will be made available through the project's website. The source code will be implemented in Python. The UNT Libraries has experience making software tools and scripts developed locally available through GitHub (https://github.com/unt-libraries).

Publications

Dissemination Through Research Publications: The new findings of the work will be published yearly at conferences and in journals. To ensure free access to publications, the PI will target high quality peer-reviewed open access journals and conferences. Pre-publication pre-prints of the papers will also be made available through the website (to the extent permitted by the copyright restrictions imposed by the publisher). Some of the venues that we plan to target include JCDL, TPDL, AAAI.

Dissemination Through Organized Workshops and Invited Talks: Workshops related to the topics of this project will be organized in top-tier conferences. The PIs have a strong track record in presenting at workshops and conferences to discuss the work being carried out during the project. Links to all workshops and conferences where this research is presented will be shared via the project website at the UNT Libraries. Additionally, any slides or presentations related to this project will be deposited in the UNT Scholarly Works Repository (https://digital.library.unt.edu/scholarlyworks) for long term access and discovery.

Appropriate Protection and Privacy

The data collected and aggregated into publicly available datasets includes data present in library catalogs and digital collections platforms that are currently open to the public for access. Web archival data used in this project likewise was collected from the state domains for Texas and Michigan by state organizations in those states. There is no data being collected or aggregated that includes private information or information collected from any protected research class and therefore does not require IRB approval.