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 research project to evaluate the use of machine learning algorithms to successfully identify and extract publications contained in existing Web archives as a way of identifying these documents to empower libraries, archives, and museums to meet their curatorial missions. This research project seeks to extend the usefulness of existing Web archive collections by designing machine learning techniques for the curation work flows of these collections aimed at discovering a way to automatically identify and extract important “publications” or “documents” from these collections. We will complete a two-track effort during the proposed project to: 1) *collect data from Web archive curators in three primary areas: a) institutional repositories; b) state and local government collections; and c) federal Web archive collections, using structured interviews and questionnaires*. This qualitative data gathering will inform the second track of work, which will: 2) *focus on the design of machine learning methods to aid in the automatic extraction of “publications” or “documents” of interest to the collection curators*. The project team includes PI Mark Phillips, Associate Dean for Digital Libraries at the UNT Libraries, Co-PI Cornelia Caragea, Assistant Professor in the UNT Computer Science and Engineering department, and Lauren Ko, the Head of the Software Development Team in the Digital Libraries Division of the UNT Libraries. Two graduate students will assist in the project. Finally, an advisory committee of at least six professionals from collecting institutions and machine learning researchers will provide guidance and advice for the project. We respectfully request $318,989 in support.

National Need

In the past decade, an increasing number of institutions across the country have incorporated Web archiving into their digital collections infrastructure as a way of collecting, preserving, and providing access to unique resources published on the Web. These collections either extend existing collections historically curated at their institution or represent a new collection area. Collecting Web content involves operating Web harvesters to programmatically spider and download large amounts of content made available through websites and Web applications. Many organizations make use of hosted services such as Archive-It, a platform operated by the Internet Archive, or use one of a variety of tools such as the Heritrix Web Crawler to collect these resources. Currently, the primary access to collected Web content involves making it available for “replay” using a tool such as Open Wayback or Python Wayback software. These replay systems typically require the user to know the URL of interest upfront and then the tool provides a mechanism for accessing the archived Web content.

Through this research, we *first* seek to better understand the tasks of a subset of Web archiving curators, specifically those that are charged with collecting, cataloging, and providing access to documents and publications for their specific organization. It is common for institutions such as university libraries and archives to archive their university domain, or for a state or local government to archive their state’s Web presence. While access to the look and feel of the Web content is important, the primary motivation is often to preserve the documents and publications that are made available to the public via the website of the organization or governmental entity. Hence, many government information professionals exist across the country who are interested in levering the large collections of Federal information, e.g., the 2008, 2012, and 2016 End of Term Presidential Web Archives, to extract and build more curated collections. *Second*, we aim at designing machine learning techniques that will automatically sift through millions of Web Archive documents and will present “just the right” digital content to the end user.

Our proposed project aligns with one of the stated IMLS goals in its strategic plan, *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. The project will explore innovative approaches to identifying and extracting publications from Web archives. These publications can then be incorporated into existing digital library platforms and further aggregated through initiatives such as the Digital Public Library of America in ways that current Web archives cannot.

Project Design

Our research project tests the assumption that we will be able to successfully apply automated classification algorithms to collections of archived Web content as a way of extracting documents and publications that are in scope of collection development policies for a variety of organizations. In order to better understand the notion of “in scope” to a collection development policy, the project team led by PI Phillips and the UNT Libraries will carry out a series of structured interviews with collection staff. In addition to interviews, our team will also gather and manually label collection development policies for these organizations as well as characteristics of existing collections of publications, such as institutional repositories, digital library collections of state and local publications, and the Technical Reports Archive and Image Library (TRAIL) that contains federal technical publications. The data manually labeled by this team will inform the automated classification portion of the project led by Caragea. The machine learning team will develop a workflow to extract novel features and design scalable machine learning techniques to extract publications from Web Archive (WARC) container files commonly used in Web archives. The extracted document features will be used to train and test a number of classification algorithms with the primary goal of being able to accurately identify documents and publications from Web archives that collection curators would identify as “in scope”. In order to train and test these classifiers, both teams will work to develop a training set of publications for the three identified collection areas for use in the experimentation. An advisory board consisting of at least six members will further aid the project by offering guidance and reviewing documentation and methods. We expect that the advisory board will be composed of a mixture of collection managers and machine learning professionals.

Outcomes and National Impact

Our primary goal is to test the precision and recall of documents identified and extracted from Web archives as being “in-scope” to an existing collection development policy in one of our existing target organizations. Other outcomes are: 1) A whitepaper describing collection development decisions and scoping from our identified collection curators; 2) 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; 3) Feature extraction workflows and methods that are identified as informative for document classification. The project teams will share all of our results via GitHub repositories maintained by the UNT Libraries at https://github.com/unt-libraries/ for code and documentation with the published output of the project made available in the UNT Scholarly Works Repository and UNT Data Repository.

Budget:

We respectfully request $318,989 in IMLS funds: $136,030 in salaries (Phillips, 10%; Caragea 2 months; Ko, 5%; 2 graduate research assistants 50%), $48,771 in benefits (15.45% federal fringe benefits rate for faculty/staff and 8.65% for graduate students), $35,648 in tuition reimbursement for the graduate students; $4,000 for travel to digital library and machine learning conferences, $2,000 for publication fees and $92,539 in indirect costs at UNT’s federally negotiated indirect cost rate (Dept. Health and Human Services, 06/19/2015).