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 Weka1 , SVMLight2 , and Mallet3 . 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 C++, Java, and 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, AAAI, WWW, ACL, EMNLP, NAACL, JAIR, TWeb, and TKDE.

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.