The University of North Texas (UNT) in partnership with the Internet Archive, and in collaboration with the PEGI Project, seeks a National Digital Infrastructures and Initiatives National Leadership Grant. UNT is requesting $249,832 in support of a two-year implementation project that will design and deliver a proof of concept tool to surface, develop, and curate collections of targeted U.S. government resources from the End of Term web archive, tailored for relevant information needs of environmental justice researchers and practitioners. This project will expand on and adapt existing technology models by facilitating the discovery and access of traditionally siloed federal agency information in web archives. The tool will be developed in consultation with environmental justice community experts and will have the goal of curating a collection of environmental data and publications critical to understanding the economic and social well-being of communities and the health and livelihoods of individuals. The outputs of the project will be designed and shared with the goal of reusability and adaptability to other web archive collections and levels of government.

The work of the project will be carried out collaboratively by Roberta Sittel (UNT, PI/Project Director) and Dr. Sawood Alam (Internet Archive, co-PI) in partnership with consultant Antoine McGrath. Additional support will be offered by Deborah Yun Caldwell, James R. Jacobs (Stanford University), Lynda Kellam (University of Pennsylvania), Shari Laster (Arizona State University), and Scott Matheson (Yale University), who make up the PEGI Project team. This project builds on the team's prior exploratory work of the Preservation of Electronic Government Information (PEGI) Project National Forum Grant [LG-88-17-0129-17] and the *Environmental Scan of Government Information and Data Preservation Efforts and Challenges*.

Project Justification

By improving access to and use of web archive collections, this project will fulfill IMLS Agency Goal 3 (Advance Collections Stewardship and Access). This project meets the objectives of NLG Program Goal 3 (Improve the ability of libraries and archives to provide broad access to and use of information and collections with an emphasis on collaboration to avoid duplication and maximize reach) in the following ways:

- Objective 3.1: Surfacing and reorganizing discrete government information resources hosted within larger sets of web archive collections will advance digital inclusion. Developing open-source technologies other groups or libraries can use or adapt addresses the need for greater access to content from archived websites.
- Objective 3.2: This project promotes innovative approaches to collection management by increasing access to born-digital content. Surfacing or extracting born-digital content into distinct and relevant collections that are designed to meet the needs of designated communities will propel subsequent digital preservation and curation activities through increased usage and other metrics that show the urgency for this work, leading to more reliable access for the long term.
- Objective 3.3: This project will examine the information needs and expectations of a specific user community so that the resulting output has an impact on their current and anticipated activities.

The project's focus is federal government information within the web archive (WARC) files generated by the End of Term (EOT) project.[1] This focus was chosen because of the aim to develop "a community of practice to preserve and provide access to electronic government information" while seeking "to build capacity for libraries to preserve historically significant born-digital government information."[2] The project team brings experience working in Federal Depository Libraries, where all have seen the shift from print to digital information delivery accompanied by a lack of responsive actions to manage and preserve digital collections with the same attention and urgency as print collections.

Since 1813, libraries have served as partners with the federal government to provide public access to government publications. Today, the Federal Depository Library Program (FDLP) accomplishes this mission in partnership with over 1,100 libraries through deposit, request, purchase, and service models; in many states, similar models are in place to promote access to state publications, though these programs differ widely depending on priorities and resources. Participating libraries also provide tools, services, and expertise that improve public access by aiding discovery and increasing context. However, to date there is no comparable, comprehensive model for born-digital government information.[3] Outside of the FDLP, publicly-accessible born-digital government information preservation has been managed primarily through unfunded volunteer web archiving activities like EOT or through triage models.[4] These projects help to preserve government information as it exists within point in time web archives but, to date, do not provide the tools needed to discover relevant and useful materials within the larger corpus of web archive files that they generate. This project does not seek to change the existing structures of the FDLP, nor the current practices in place for capturing born-digital government information. Instead the team will work with EOT content as an example of an existing web archiving project, to explore methods that increase discovery and access to this information. The model developed by this project will be applicable to government information produced by other jurisdictions, meaning that libraries and other cultural heritage institutions can adapt the resulting tool for the needs of other designated communities.

In its 2018 National Forum, the PEGI Project explored the needs of a broad stakeholder community for the preservation of born-digital government information.[5] Forum findings identified challenges related to digital preservation that included disjunctive incentives to cooperate, an overly broad landscape of need, and a lack of tools calibrated to the particularities of content generated by government entities. Additional findings noted the need for "solutions that closely align with public needs, making sure they are both diverse and inclusive." This project will respond to these observations by scoping our work to a specific topic, Environmental Justice, and by engaging directly with groups working in EJ activism and research to better understand particular information needs. The team will

[1] Established in 2008, the End of Term (EOT) archive is a collaborative project between academic, government, and public institutions charged with capturing the entire .gov and .mil domain at the conclusion of each presidential term. Every four years, the amount of data captured by the EOT grows exponentially; for example, in 2012 EOT crawled 3,273 websites and in 2016 the project crawled 53,324 websites. Captures of this content offer a wealth of resources for researchers and policymakers but getting to it in a relevant or deliberate way is not currently possible. End of Term project partners for 2020 included the Internet Archive, University of North Texas, Library of Congress, Stanford University, Government Publishing Office (GPO), National Archives and Records Administration (NARA), and Environmental Data and Governance Initiative (EDGI). https://eotarchive.org

[2] https://www.pegiproject.org/objectives

[3] GPO's **govinfo** database is an ISO 16363 trustworthy digital repository, but contains primarily Congressional information with only a very small amount of historic publications from the Judicial and Executive branches.

[4] See "Planning a Community-Created Data Rescue Toolkit" [LG-72-18-0102-18].

[5] See "PEGI Project National Forum Summary and Report" and the two-year report "Toward a Shared Agenda," available from https://www.pegiproject.org/publications.

utilize existing mechanics, metadata, and tools for identifying and discovering government information to help surface resources relevant to this project. This project will create tools that leverage content already collected through cooperative activity and disseminate the resulting products for integration into future platforms.

For the Surfacing Digital Government Information project, the large collection of EOT materials[6] will be scoped to environment- and sustainability-focused information. This focus aligns with the Biden Administration's stated goal to advance environmental justice across the federal government.[7] The breadth of need for discovery and access to government-published resources includes environmental non-governmental and public advocacy organizations, Indigenous communities, legal professionals working on behalf of public interest groups, and community members who are impacted by the legacy of environmental inequities. Federal agencies that disseminate information resources pertinent to their research include EPA, NOAA, Interior, Commerce, Defense, HUD, HHS, and many more. Researchers seek public information that is now distributed solely on the web and siloed within separate government domains (epa.gov, doi.gov, energy.gov, etc.), to conduct due diligence, dispute resolution, policy-making, planning, and community activities. However, the internet is not a fixed medium, and organizations, including government entities, remove or replace content on a regular basis. While programs such as the FDLP, and other government initiatives, have long afforded publicly-accessible preservation for print materials, similar models for born-digital content are limited by collection priorities.[8] As a result, the content that researchers need may no longer be on a public website, necessitating exploration of preserved web archives.

Web archiving mitigates loss of information from websites. With an eye to current and future research and information needs, projects like the EOT were developed to capture the whole of the government web domain at the end of each presidential term. Regular, systematic, web archiving is a valuable service that captures an organization's public-facing information at a given point in time. However, it is difficult for members of the public to find and use content within web archives as they currently exist. Information seekers typically rely on discovery tools to find what is needed, but these types of search and discovery tools largely do not exist for web archives.[9] Individual websites captured in web archives may be cataloged, but individual publications embedded within the web archives are neither cataloged nor made easily discoverable. Additionally, these materials are not holistically developed or curated as collections, which can make it a challenge for novice researchers to know where to start looking.

The Internet Archive (IA) has long been a vital partner for governments, universities, and cultural heritage organizations to gather born-digital materials.[10] Despite the ever-growing cache of web archives,

[6] EOT 2020 collected approximately 300TB of data. Four years earlier, EOT 2016 collected approximately 200TB of data (which included FTP sites for the first time and therefore greatly expanded the amount of data collected). The full scope of EOT collections comprises approximately 550TB of data.

[7] https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/26/fact-sheet-a-year-advancing-environmental-justice/

[8] See "What's available?" at govinfo.gov https://www.govinfo.gov/help/whats-available and "About Congress.gov" https://www.congress.gov/about

[9] Milligan, I. (2016). Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *International Journal of Humanities and Arts Computing*, 10(1), 78–94. https://doi.org/10.3366/ijhac.2016.0161

[10] The Internet Archive hosts content including digitized resources, broad web crawls that can be explored using its Wayback Machine platform, and targeted crawls captured by libraries and cultural heritage institutions the Archive-It subscription service. For example, the Government Publishing Office (GPO), the California State Library, and other official US and international institutions contract with Internet Archive to archive government websites.

broad-reaching solutions for making these collections accessible and useful have not materialized.[11] Research exploring the usefulness of web archives ranges from user testing models[12] to corpus analysis[13] and exploratory efforts intended to extract content or metadata based on topic, timeframe,[14] or filetype.[15]

To further complicate access to government information, current access systems do not provide the necessary context, provenance, or relevant details needed to interpret and use government information. For example, placing a document in context often requires that a user know the originating agency, the regulatory underpinnings, and the reason it was created. Interpretation of government information resources depends on this context and other relevant details that are artifacts of its production and distribution. In print collection management practices, libraries have robust historical systems for identifying and describing government information to provide this context and aid in discovery and continuity, but they are not applied to documents hidden in web archives. To use tools and techniques that aid in discovery and navigation of content, libraries need tools to extract and describe the content so that it can be made available in a user-friendly form.

The Surfacing Digital Government Information project innovates on prior and current work aimed at creating models and technologies to make content from web archives more discoverable and useful. Recent projects have seen success in using existing metadata, including DOIs, as well as machine learning models, to identify and surface academic articles and other publications from web archives.[16] Where prior and current projects have had success, these practices have been only narrowly applied to government information, if at all.[17] Government information carries a unique set of problems and expectations compared to academic literature. For example, government information on the web is primarily structured and presented for its immediate intended purposes and audiences, and lacks features that would facilitate future discovery, transformation, and analysis. Unlike academic literature, it is rare if a citation to a government information resource provides all of the information necessary to identify and access the source. By designing tools that take advantage of features specific to government information in order to better leverage emerging discovery and access technologies, this project will directly address several challenges surfaced in the PEGI Project National Forum, as discussed above.

Project Work Plan

This implementation project will run from August 1, 2022 to July 31, 2024 and will consist of three interrelated project parts. The first, which will span the project lifecycle, is to develop and refine a set of tools to identify and surface government information from archived federal websites based on

[11] The Internet Archive is now testing a beta version of a search tool for content within archived websites. While it is helpful for tracking down known materials, its usefulness for discovering content relevant to a topic or theme is limited.

[12] Abrams, S., Antracoli, A., Appel, R., Caust-Ellenbogen, C., Denison, S., Duncan, S., & Ramsay, S. (2019). Sowing the Seeds for More Usable Web Archives: A Usability Study of Archive-It. *The American Archivist*, *82*(2), 440–469. https://doi.org/10.17723/aarc-82-02-19

[13] Archives Unleashed. https://archivesunleashed.org

[14] Gossen, G., Risse, T., Demidova, E., (2020). Towards extracting event-centric collections from Web archives. *International Journal on Digital Libraries*, *21*(1), 31–45. https://doi.org/http://dx.doi.org/10.1007/s00799-018-0258-6

[15] Tarver, H. and Phillips, M.E. (2018). Labeled PDF Dataset from Texas Records and Information Locator (TRAIL) Web Archive. https://digital.library.unt.edu/ark:/67531/metadc1757660/

[16] This project takes as one of its models the recent development and launch of Internet Archive Scholar. See: https://scholar.archive.org/.

[17] See UNT's IMLS project "Programmatic Extraction of 'Documents' from Web Archives" [LG-71-17-0202-17].

scoped parameters. The second is to collaborate with partners to identify types and sources of government information that support communities' organizing and research efforts associated with environmental justice (EJ). The third is to contract with a library professional to conduct a scan and needs assessment focused on access to environmental justice resources. The project will use an iterative approach to refine the tool based on findings from the team's collaboration with EJ communities and researchers and outcomes from the scan and needs assessment.

Part 1: Web Archive Content Surfacing Tool

The first part of the project will result in the development and refinement of a "surfacing" tool that will allow for the discovery of targeted U.S. government-created information from websites in the End of Term archive (EOT). To develop the surfacing tool, the project team will employ technology consultant Antoine McGrath, whose prior work includes developing the code behind CRSreports.com, which was a site that collected Congressional Research Service (CRS) reports from thousands of disparate websites and generated accompanying metadata. For the Surfacing Digital Government Information project, Mr. McGrath will develop and oversee the technical plans necessary to achieve the project's goal of implementing a proof-of-concept tool that will identify target materials within archived web content that can then be organized into discrete, relevant, and easily discoverable collections.

Roberta Sittel, Project Director, in consultation with Mr. McGrath and co-PI Dr. Sawood Alam, will hire a UNT graduate student to work 20 hours per week for the duration of the project. The student will conduct scripting, database infrastructure design and creation, and iterative output testing. Mr. McGrath will provide guidance and training, set expectations, and review the work of the graduate student. Additionally, Mr. McGrath will coordinate with Dr. Alam to ensure appropriate permissions are in place for access to IA infrastructure and Web Archives (WARC) files from the EOT. Dr. Alam will also assist Mr. McGrath with access to a researcher virtual machine (VM) environment, which will enable testing the harvester and classifier tools throughout the development process. Dr. Alam will meet monthly, or as needed, with Mr. McGrath and other project team members to assist with project planning, implementation, and assessment to ensure access to the related data is available for the project's success. During the development and refinement phases of the project, Mr. McGrath, the student, and the project PI will also meet weekly, at minimum, to ensure the work is on schedule and that the student has the information and resources necessary for success.

The Web Archive Content Surfacing Tool (technical diagram) will consist of two components, a harvester and classifier. The harvester will extract and locally store filetype HTML/PDF documents from the EOT WARC files. A filetype identifier will be used and the HTML/PDF will be stored as its fingerprint sha256 which will prevent exact duplicates. Processed content from the WARC files will be noted in a Database (DB) for exclusion as will its resulting sha256, allowing later analysis of content hot spots as well as trends in duplicate sourcing of exact files. The classifier will evolve throughout this project. Initially the classifier will rely on expected content discovery within manually defined bounding boxes. In later phases we will utilize Natural Language Processing (NLP) to determine target match, and eventually we will test image classifiers. Perceptual hashing is an additional option for content deduplication which can be explored as a means of improving process efficiency. Similarly the SHA1 of HTTP responses may be utilized to avoid reanalyzing unchanged content.

The classifier will be refined based on findings from the other two parts of the project (detailed below), so that the collection created reflects the needs of community stakeholders. The stored metadata will then be amenable to analysis, and front end search interfaces can be made to serve select types of content based on sought terms. This model is based on the Internet Archive's successful development of Internet Archive Scholar, a cutting-edge discovery tool created by harvesting scholarly literature via DOIs.

The Web Archive Content Surfacing Tool will identify sought content within WARC files, which will then be collected, analyzed, and programmatically evaluated with respect to the intended parameters for inclusion. The result will be more easily discoverable content. Using currently available technology we are able to identify publications that might not otherwise be discovered by casual researchers who are not able to programmatically parse WARC files. The result is therefore greater access to content embedded within web archives, and the theme of the content can be adapted to additional file types or collections.

Concurrently, the project team, all subject matter experts in government information, will identify and scope the initial parameters for the tool. Team members will use data generated by the project's initial outputs to identify gaps within existing metadata that impede the discovery of content embedded within web archives and digital collections. Metadata analysis addresses one of the thematic concerns that emerged during the PEGI Project forums.[18]

The project team and consultants will undertake an agile software development framework employing a "build with not for"[19] civic engagement focus for the surfacing tool across the two years of the project, and release the tool as open-source code. Agile development involves frequent user feedback and continual improvement to the software through many iterations to ensure the software matches the users' needs. The project team's engagement with stakeholders working on environmental justice issues will inform the refinement of the tool, using methods such as open-ended interviews to elicit needs, and review of output to validate results, in order to improve the identification and classification of content based on scoped parameters. Findings from these engagements will be evaluated for integration into refining the tool's parameters and functionality as the development progresses. Developing the surfacing tool to serve a focused set of needs related to environmental justice research creates a manageable scope for this project while allowing the final product to have the capability to be modified to a similarly focused experience for other designated communities.

The completed version of the code for the Web Archive Content Surfacing Tool will be made available as an open-source project through GitHub and backed up following standard practices for providing access to project data through an institutional repository. Outputs from the iterative testing and development of the surfacing tool can be made available as searchable IA collections for public use. Outputs will be assessed throughout the iterative development process. These will be measured as the number of WARCs evaluated, number of unique files processed, documents identified, review of the document classifier criteria, and analyzing output documents to determine both the extent of their match with the criteria as well as their usefulness to the consulting practitioners. As part of its findings, the project team will also develop functional requirements for a permanent implementation of this tool.

While the team included delivery of a proof of concept interface in the preliminary proposal, the scope and timeframe of the project will not allow for this. Instead, the project will host a collection at Archive.org that will provide a testing interface for team members and external partners. This collection will facilitate exploration of basic search strategies and functionality for the content surfaced by the tool. From this collection, the team will identify gaps within the existing metadata that impede discovery of content embedded within web archives and digital collections. The project will utilize public access points and existing metadata at the Internet Archive to conduct a gap analysis between existing metadata sources and those required to meet user needs to work across digital collections.

Part 2: Environmental Justice Community Engagement

The second component of the project will engage partners working on environmental justice issues to identify types and sources of government information that support their research and work in

[18] PEGI Project National Forum Summary and Report [#LG-88-17-0129- 17] https://www.pegiproject.org/publications

[19] https://buildwithnotfor.wordpress.com

addressing problems in their communities of practice. By engaging directly with user communities, the Surfacing Digital Government Information project will build a collection and set of tools with users.

The team will collaborate with partners working on environmental justice efforts with communities of color and economically marginalized communities, and academic researchers in related fields. In order to address the information needed by community stakeholders, researchers, and policy-makers, the project team will work with individuals and groups who are already active in these areas and who have committed time and resources to build relationships and understand the needs of marginalized communities. A two-year project does not allow sufficient time to successfully engage stakeholders in relationship-building. As a new entrant to the space of environmental justice, the project team will focus on relationship development, following appropriate models for academic and community partnerships. The project team will work with partners as described in the Diversity Plan to identify barriers to information access that are critical to decision-making for environmental justice advocacy and policy.

Project Director Roberta Sittel will hire and supervise a social science graduate research assistant to assist the team in developing instruments for interviewing practitioners and conducting focus groups. Following appropriate protocols for human subject research, the graduate assistant and the team will host focus groups to identify specific information needs that will inform the requirements for the classifying and extracting tool. When the test collection interface is available, testing will be conducted with some of these same practitioners using web-based tools and other relevant methods to gather information needed to refine the output of the surfacing tool. This part of the project will aid in determining how different user groups approach searches and determine document needs. These research activities will identify the metadata elements necessary to support these information-seeking behaviors. This insight will enhance the project's development and refinement of the surfacing tool leading to improved discovery and access of relevant government information for localized use. A final round of semi-structured interviews and focus groups will be conducted to identify remaining information needs and discovery problems.

Work with these community partners will be assessed through qualitative and participatory feedback, as well as quantitative feedback on test search results that surface relevant results coupled with the number of issues identified. Additional measures will include data from the web-based tools employed for engaging these stakeholders and from identified missing metadata elements. This feedback will inform the iterative development of the tools.

Part 3: Environmental Scan/Needs Assessment

In the third part, conducted concurrently with partner engagement, the project team will contract with a library professional with expertise related to marginalized communities in order to conduct a scan and needs assessment focused on access to environmental justice resources. The goals are both to refine the development of the proof-of-concept surfacing tool and to inform the functional design and needs for an interface to implement the tool.

This scan and needs assessment will focus on government information resources in use for research and practices related to environmental justice. The contracted library professional will review available literature and resources, and then collaborate with project partners, including the PEGI Advisory Board,[20] EDGI, YCEJ, and faculty partners to design and conduct a focused scan and needs assessment. The resulting output will be published as a report and used to inform iterative design and testing practices for subsequent applications for this tool and similar projects. The output for this part of the project will be a report published at www.pegiproject.org.

[20] The PEGI Project Advisory Board includes: Heather Christenson, Gregory Eow, Ed Garcia, Debbie Rabina, Vicky Reich, Daniel Schuman, Michelle Trumbo. https://www.pegiproject.org/project-team/#advisoryboard

To facilitate sustained, broad dissemination of all outcomes and digital products produced by this project, the team is requesting funding for web hosting, domain retainment, and other related needs. During the two-year project, the project team will host quarterly webinars targeted toward library and information professionals to disseminate findings and build engagement around government information digital collection development. The team is committed to submitting proposals to relevant conferences and academic journals where the findings of the project can be shared with broader audiences. The team will also post project updates and initial findings through the project's lifecycle to the blog hosted at www.pegiproject.org.

Diversity Plan

The environmental justice (EJ) movement emerged in the 1960s alongside the civil rights movement, meaning it is explicitly an issue of diversity, equity, and inclusion. The environmental justice movement seeks to acknowledge the disproportionate impact of environmental harms on marginalized communities and respond to these harms through research, advocacy, and policy.[21] EJ work may be conducted by community-based groups, researchers working in partnership with community stakeholders, and other groups seeking solutions to social issues that intersect with the environment. To better understand how resources in web archived collections can be used in seeking solutions to complex issues, this project adopts a thematic focus of environmental justice (EJ)[22] and will engage directly with researchers and practitioners to better understand user needs for access.

There is often an inherent divide between the researchers and practitioners working in communities to solve complex problems, including environmental research and EJ communities. Similar divides extend into libraries and archives despite our expressed intentions to serve the public and meet access needs. Historically there is an absence of practices for connecting directly with users to develop more targeted approaches for meeting information needs. By working with a specific community of practice, EJ, and scoping to information that can inform their work and research, this project will offer a framework for libraries and archives to engage more deliberately with specific user groups to better understand information needs and inform collection development practices.

Recognizing the challenges of building meaningful working relationships in the short period of time that the project will comprise, the project team will collaborate with partners who are currently working with communities of color and economically marginalized communities. The project team defines marginalized communities as those that have historically been excluded from civic participation and decision making. Knowing that these communities vary from place to place, the contracting professional will draw from the work of policy researchers, community organizers, and environmental justice non-profits, to produce a report that will contribute to an understanding of gaps in access and inform urgently-needed work. Recruitment of participants within marginalized communities can be difficult and sometimes seen as problematic or exploitative precisely because of the marginalized nature of the participant. The work of part 2 of the project involves engaging with the relevant communities in a respectful and transparent way to determine the challenges they face and the opportunities the project's activities can facilitate improvement to information access. The opportunities will also inform future practice in curating community-focused collections from existing and future web archives. The project team will work with the partners identified below to build the connections needed to engage in a

[21] For a brief overview of the history of EJ within the US, see: Perez, Grafton, B., Mohai, P., Hardin, R., Hintzen, K., & Orvis, S. (2015). Evolution of the environmental justice movement: activism, formalization and differentiation. *Environmental Research Letters*, 10(10), 105002. https://doi.org/10.1088/1748-9326/10/10/105002
[22] See https://www.epa.gov/environmentaljustice and https://www.doi.gov/blog/agenda-rooted-environmental-justice

way that addresses community needs that are historically overlooked or dismissed. The project team is committed to building further connections following the project launch and fairly compensating participants for their time and expertise.

By focusing on environmental and sustainability information resources in use within communities for their research, the project will seek and identify solutions to information access barriers. Outputs will contribute to collections that support decision-making within and relevant to representatives from community partners. The project team will also contract with a library professional with research expertise in order to conduct a scan and needs assessment focused on access to environmental justice resources.

Since 2016, the Environmental Data & Governance Initiative (EDGI) has worked with "organizations and communities concerned about climate change, science policy, good governance, and environmental and data justice."[23] By partnering with EDGI, the project team will be able to consult with experts on methods for identifying gaps in access to environmental data and information. Through structured meetings with EDGI staff and volunteers, the project team will draw on EDGI's established and ongoing work. EDGI's Environmental Data Justice Working Group assesses existing models for data infrastructure, storage, and dissemination with the goal of creating collaborative, equitable, and transparent data practices and civic technologies.[24] EDGI's work will assist the project team in identifying information needs that can inform refinement of the surfacing tool. By generating a tool that will supplement EDGI's efforts, this relationship is intentionally mutually supportive as their work will assist the project team with identifying needs within specific communities and ours will expand access to existing stores of information.

EDGI will also work with the project team to identify community groups that are interested in contributing to project outputs as described in the work plan. Community experts, participating community organizations, and practitioners will be appropriately compensated for their time and expertise.

The project team will also partner with the Yale Center for Environmental Justice (YCEJ) to identify environmental justice practitioners and researchers. Professor Gerald Torres and other YCEJ faculty will connect the project team to community groups and scholarly communities working in environmental justice in order to learn more about their information needs and how the output from the surfacing tool may be improved to serve these needs (via surveys and/or focus groups).

To better understand the needs of academic researchers and those that support undergraduate and graduate students, the project team will work with Dr. Alexandra Ponette-Gonzalez, Associate Professor in UNT's Department of Geography and the Environment. Dr. Ponette is one of seven members of the EPA's Clean Air Scientific Advisory Committee (CASAC) and her research explores the "interactions between humans, the atmosphere, and the biosphere in the context of a changing environment."[25] Through this partnership, the project team will engage with Dr. Ponette's academic circle of students, peer faculty, and researchers to better assess the information needs in support of the communities with whom they work.

Through the work with these partners, the project will demonstrate community needs led development. Tool modification and content served will be based on input from communities consulted, especially those under-represented in current systems designs. The transparent development process that we are committed to allows for others communities and individuals to copy the project for their own focus.

[23] https://envirodatagov.org/about/
[24] https://envirodatagov.org/environmental-data-justice/
[25] https://ponettelab.cargo.site/

Project Results

This implementation project will create a set of tools that advance innovative approaches to digital collection management and discovery. The tools will enable users to take advantage of the metadata embedded in archived websites to surface, explore, discover, and curate, born-digital government information resources that have been captured using widely-available web archiving technology. The developed tools can be adapted for extracting content from archived websites from outside of the federal government and further refined through collaborative work with researchers and practitioners. Libraries and cultural heritage institutions can apply these tools to create more meaningful collections from masses of web archives. This project represents a step toward the development of a digital government information library by leveraging metadata specific to government-produced content. This project will also produce publicly-available, focused collections tailored to resources that are relevant to environmental justice research, and a lifecycle-informed needs assessment for future library work in support of environmental justice. Finally, this project will add to the body of professional expertise about approaches to harmonize digital collection development with current and anticipated community research needs.

One key result from this project is to create a better way to analyze archived websites and extract relevant content for use by employing NLP fine-tuned for a specific content use case. By focusing on government information, this implementation project will engage with users and analyze existing metadata schema to better understand how NLP can enhance discovery of materials and allow for curation of collections from web archives. This tool will also fit a stated need to add to available resources for both library collection development and researcher use that are directly tailored to work with government information resources.

The resulting code will be made available in the PEGI Project's GitHub repository and in other digital repositories for librarians, digital archivists, and researchers to use and adapt to other web archives. Iterative versions will be shared along with relevant lessons via the pegiproject.org blog to demonstrate how the tools were reviewed and refined throughout the project's lifecycle.

By making captured content more usable, this project will create added incentive for digital preservation strategies that rely on web archiving, and surface unique materials within the Internet Archive's End of Term collection. The model and tools developed by this project will advance the long term goal of improving access to the universe of digitized and web archived government content to advance the previously mentioned goal of developing a digital government information library. To advance future work leading toward this ambitious goal, the project team will produce functional requirements for infrastructure, including a platform, interface, and service model, that would facilitate use of the tool.

Research engaged with practitioners and experts will improve the functionality of the tool and generate relevant data about how researchers classify, consider, and use government information resources within their field of work. Stakeholder engagement will drive the creation of usable collections hosted on Archive.org, and inform the functional requirements for platform implementation as well as guidance created to accompany the code for those wishing to implement the tool in their own computational environment. By working with practitioners, the project outcomes will impact federal agencies, like the EPA and Interior, by improving access to the agency's archived content across the federal web domain.

The work of the contracted librarian researcher will complement the work with the stakeholder communities and add to the growing body of research around usability of government information sources, including archived web content. These outcomes will be published as a needs assessment, adding to available literature that refocuses government information discovery from a producer to a user perspective.[26]

During the two-year period, the project team will host quarterly virtual presentations developed for a broad library and information science community, to share interim products and findings. Attention will be given both to the technical aspects of the tool development and refinement, and to the process of collaborating with communities of practice, conducting productive focus group research in partnership with community experts and researchers, and learning from practitioners how libraries can better serve their needs. Throughout the grant period, the team will seek opportunities to present at relevant national and international conferences as well as publishing in appropriate academic and professional journals. At the end of the project, the team will publish a report that will outline work accomplished, outcomes, and challenges encountered.

[26] For an example of prior work from the PEGI Project, see: Lippincott, Sarah K. *Environmental Scan of Government Information and Data Preservation Efforts and Challenges*. Atlanta, Georgia: Educopia Institute, 2018.