**Environmental Impact Statements: Public Archive Project**

**Northwestern University**

**INTRODUCTION:**

**GRANT SUMMARY:**

Northwestern University requests $500,000 from the Andrew W. Mellon Foundation to transform the world's largest archive of U.S. Environmental Impact Statements (EIS) into a globally accessible digital resource. Our EIS collection is the authoritative collection to which the EPA directs researchers. These documents record policy history by uniquely capturing the rhetorical debates between federal agencies, corporate interests, and the impacted communities, offering a rare window into how environmental decisions have been defended, contested, and reimagined over time. This ambitious project will complete the digitization of **[NUMBER]** thousands of remaining documents, build a discovery platform styled as a public exhibit, and develop AI tools that unlock the collection's potential for new forms of environmental and humanistic inquiry.

**REASON FOR THE WORK:**

**A Unique Historical Archive with Urgent Contemporary Relevance**

Environmental Impact Statements, required by the National Environmental Policy Act of 1969 (NEPA), document the predicted ecological, social, and economic consequences of major federal actions as well as considering impacts of alternatives to those actions. With recent efforts to weaken NEPA implementation, this archive can serve as a vital benchmark for understanding how environmental decisions have been justified, and how those justifications may shift in the future. Northwestern's extensive collection began with a 1992 donation of over 20,000 EIS titles from Northwestern faculty H. Paul Friesema, a leader in environmental research. Through systematic collection efforts, this archive has grown to become the authoritative record of U.S. environmental assessment, offering a rare longitudinal lens into how environmental decisions have been made, rationalized, and contested at the federal level.

As climate change intensifies and environmental justice rises in urgency as a defining issue, the EIS archive represents a critical tool for historical research into state power, environmental governance, and the cultural narratives that have justified ecological transformation. In particular, the community response sections within these documents stage a complex and often overlooked debate between federal agencies, industry lobbyists, and the communities most directly affected by proposed environmental interventions. This archive enables scholars to compare the rhetorical positions of these three constituencies over time, across regions, and within varied ecological contexts, aiming to surface the hidden voices of impacted communities as a chorus that challenges and critiques the dominant logics of industrial development and governmental authority.

Despite the collection's immense potential, limitations currently constrain its impact. While Northwestern has successfully digitized over **[33,000]** EIS documents through a partnership with Google, making them available through HathiTrust, in-depth research using the remainder of this collection remains accessible only through physical visits to Northwestern's library. This geographical barrier significantly limits its scholarly impact and utility for addressing urgent environmental challenges. Limitations to access also include discovery of materials, as limited metadata does not support searching by key criteria such as geographic location or project type.

The proposed project directly aligns with the Mellon Foundation's mission to support work that democratizes access to knowledge, illuminates histories of inequity, and enables public scholarship grounded in primary sources. By removing geographic and technical barriers to accessing critical environmental records, this initiative will foster global collaboration on environmental research and policy development while supporting communities seeking to understand and address environmental injustices.

**GOALS:**

**Comprehensive Digital Transformation**

***Making Distinctive Local Collections Universally Accessible***

This initiative goes beyond the simple digitization of a collection. More generally, it marks the Northwestern Library's first concerted effort to publish a distinctive local collection as a dataset for broad public access and use. We aim to leverage our unique library holdings and provide new opportunities for research and education by making collections adapted a dataset interrogable in numerous ways with an interactive website, with artificial intelligence-driven tools for search, document and image retrieval, and for surfacing patterns across the large dataset with maps, graphs, and other visual modes of user engagement. It represents a new approach to collection access and use, reframing the archive as a platform for both computational analysis and public engagement.

The project encompasses four interconnected components designed to maximize the collection's scholarly and public impact. The first phase, already underway as a strategic priority for Northwestern University Libraries, focuses on digitization completion, continuing Northwestern's substantial previous investment in a specialized $150,000 scanner and established digitization workflows. There are an estimated **[34,727]** individual sheets of microfiche in the collection. This effort will digitize the remaining documents with rigorous quality control, implement OCR processing optimized for the technical documents, maps, and diagrams characteristic of EIS materials, and develop comprehensive metadata schema capturing geographic, temporal, and agency information. Crucially, this phase will preserve oversized materials such as maps and technical drawings that are frequently omitted in standard digitization efforts but contain important environmental data.

***Enabling Access through a Public Web Platform***

In phase two, the project will create an exhibit-style discovery website that transforms how users interact with this vast collection. Built on Northwestern's existing Canopy platform, an innovative library-led initiative for public exhibits and visual storytelling, this public-facing digital interface will go far beyond a searchable database. It will offer intuitive discovery tools for researchers at all levels, feature curated exhibits highlighting key environmental cases and their historical significance, and incorporate interactive timelines and visualizations of environmental policy evolution. The platform will also support community engagement through commenting, tagging, and contribution features, while ensuring accessibility compliance for diverse users.

***Unlocking Environmental Insight with AI Tools***

The project's most innovative component involves expanding the AI tools Northwestern has already developed with a previous IMLS grant to create environmental research tools that will unlock analytical possibilities for this large collection that exceeds the reading capacity of a single researcher. These tools will enable text mining to uncover patterns in federal reasoning and rhetoric, industry lobbying language, and community responses documented in EIS records, such as town hall transcripts, written letters, and public testimony. Computer vision will extract data from legacy maps and schematics, and geospatial analysis will link projects to demographic and ecological impacted areas that have not been adequately studied. Multi-modal and temporal analysis will integrate these voices and formats over time and space, allowing researchers to compare the rhetorical and policy positions of government, industry, and affected communities across historical periods and geographic regions. By enabling systematic comparison across these rhetorical positions, the platform will allow scholars to trace the evolution of environmental conflict as a deeply human, often regionally specific negotiation between authority, economic pressure, and lived experience. Our humanistic vision for AI centers on leveraging models and agents for the public good, enhancing access, discovery, and knowledge creation through an intuitive, chat-based interface that makes the archive usable not just for technical scholars but for all.

***Benchmarking Environmentally Sustainable Infrastructure***

Recognizing the environmental implications of digital infrastructure, the project incorporates environmental sustainability as a core design principle. The team will develop a carbon-aware architecture for AI computation, document and share best practices for environmentally sustainable digital humanities, and produce a detailed case study documenting the carbon costs and sustainability decisions made throughout the project. This component will serve as a model for environmentally responsible digital humanities practice, ensuring the project's environmental benefits far outweigh its computational footprint.

***Impact and Innovation***

This initiative promises to advance multiple fields while addressing urgent contemporary challenges. By democratizing environmental knowledge, the project will remove barriers that have historically limited access to critical environmental data, fostering expanded global collaboration on environmental research and policy development.

The project represents a significant methodological advancement in environmental and digital humanities. The AI tools developed for this archive will allow scholars to trace the evolution of environmental reasoning, rhetoric, and representation across five decades of U.S. governance. These innovations will set new standards for large-scale rhetorical analysis of contested public discourse, with potential applications across archival humanities, civic data research, and public history.

Perhaps most importantly, the project embodies the public humanities mission by making visible the historical patterns of environmental decision-making that have shaped contemporary inequities. Researchers, activists, and affected communities will gain access to a valuable resource for understanding how environmental harm has been systematically justified and imposed, particularly in marginalized communities. The exhibit-style interface and community features will enable new forms of collaborative scholarship and public engagement.

The project will also create a replicable global model for environmental archives. Its metadata standards, digital infrastructure, and analytical workflows will serve as a framework for environmental document preservation, discovery, and analysis in other national and international contexts.

**ACTIVITIES:**

**Multi-Stream Development Architecture**

In lieu of a conventional project timeline with stepwise milestones, our project employs a parallel development approach with five interconnected work streams designed to eliminate single points of failure while enabling continuous iteration and refinement. Throughout, our philosophy focuses on integrating librarians, faculty partners, and both undergraduate and PhD students into project teams throughout the grant.

**Phase 1: Data Foundation Stream (Months 1-24)**

Northwestern University Libraries will provide in-kind support for the ongoing digitization of the remaining EIS collection using our specialized $150,000 scanner and established workflows. The grant will support the critical post-processing work that transforms these scans into an AI-ready research dataset. Dan Zellner and Nicole Finzer will lead the continuous workflow of OCR processing, quality control, and metadata enhancement specifically optimized for both public access and computational analysis.

Aerith Netzer (Digital Publishing Librarian) will significantly reduce the costs of this phase by developing in-house postprocessing optimizations using computer vision modeling rather than relying on expensive out-of-the-box software solutions. This approach will create custom computer vision models specifically trained on the unique challenges of EIS documents, which include technical diagrams, oversized maps, handwritten marginalia, and complex multi-column layouts, while building internal capacity for ongoing optimization and refinement. This innovation not only reduces project costs but also creates more effective processing capabilities tailored to our specific archive.

This stream operates continuously throughout the project, with processed documents flowing immediately into both the public platform and AI development streams. The team will implement metadata creation that supports both human discovery and machine learning applications, including comprehensive geographic, temporal, and agency information. Rather than waiting for complete digitization, this approach enables iterative development and immediate user feedback integration. The machine-readable dataset optimized for AI modeling delivered from this phase establishes the foundation for all subsequent project components.

**Phase 2: Public Platform Stream (Months 1-24)**

The public-facing platform development divides into two distinct phases reflecting different development priorities. During months 1-12, Rachel Cole (Curator of Transportation Library), Basia Kapolka (Digital Humanities Librarian), Mech Frazier (Geospatial Librarian), and Aerith Netzer (Digital Publishing Librarian) will focus on building the exhibit-style website using Northwestern's Canopy platform, integrating geospatial capabilities, and developing visualization tools that make complex environmental data accessible to diverse audiences. Rather than creating a traditional database, this team will craft an experience that guides users through curated environmental narratives while maintaining the flexibility for independent exploration across multiple research domains.

A dedicated Knight Lab studio of 18 senior students from majors across the university will work on this platform development as a deep, year-long project that will serve as the capstone of their senior Northwestern careers, in a team-based senior thesis. This studio will focus on user experience design, interface development, and accessibility features that ensure the platform serves diverse research communities from academic scholars to community activists. The students will work directly with journalism and library partners to understand how different user groups interact with environmental archives, creating innovative solutions for discovery, visualization, and engagement. James Lee will teach this Knight Lab studio in collaboration with Jeremy Gilbert, ensuring that the platform development remains aligned with the project's humanistic research goals and public engagement priorities.

This development work will be enhanced by recruiting a Media, Publics and Rhetorics PhD candidate whose dissertation research will focus on public access website design and community engagement strategies. James Lee will serve as dissertation advisor for this student, whose embedded research will ensure that platform development remains grounded in rhetorical theory and public humanities principles while generating original scholarship about digital access and environmental communication.

Months 13-24 shift focus to intensive user engagement and testing, led by Elizabeth Shogren and her Medill investigative journalism students based in the Medill Washington DC Bureau. This field-testing phase will involve real-world investigative journalism using the EIS public access platform, community partnerships, and iterative interface refinement based on how practitioners actually use environmental archives for investigative reporting, advocacy, and policy analysis. This approach ensures the platform serves both academic research and practical applications in environmental journalism and activism, producing a public-facing exhibit website that demonstrates new possibilities for open access to environmental policy archives.

**Phase 3: AI Agent Stream (Months 1-24)**

This phase represents a methodological breakthrough in digital humanities and information science applications of artificial intelligence. Working with Han Liu (Computer Science), a leader in AI research, our team will implement the novel context engineering and Model Context Protocol (MCP) frameworks that have yet to be adopted at scale in academic contexts. This collaboration enables a fundamental shift from the massive generalist, non-specific AI models that dominate current applications toward deep, contextually sensitive systems designed specifically for library collections.

**Context engineering** is an emerging AI methodology that moves beyond the "prompt engineering" approaches popular in commercial and academic applications of AI models. Rather than trying to coax general-purpose models like ChatGPT or Claude to understand specialized domains through better prompts, context engineering systematically builds domain-specific knowledge directly into the AI's operational framework. For our EIS archive, this means training the system not just on a massive corpus of environmental documents, but on the specific rhetorical patterns, institutional relationships, legal frameworks, geographical dimensions and historical contexts that shape how federal agencies, corporations, and communities communicate about environmental issues. The contextually-honed model learns that when a document mentions "significant impact," it carries different weight depending on whether it appears in an agency assessment, industry comment, or community response, which represents a set of distinctions that generalist models cannot reliably make.

**Model Context Protocol (MCP),** developed by Anthropic but not yet widely applied in academic settings, provides the technical infrastructure that makes context engineering scalable and sustainable. MCP enables AI systems to connect dynamically with specialized knowledge sources, databases, and analytical tools rather than trying to contain all knowledge within a single large model that contains the undifferentiated content of an entire corpus. In our implementation, MCP will allow the AI agent to simultaneously access the EIS text corpus, geographic databases, scientific literature about environmental impacts, and legal precedent databases, creating responses that integrate multiple forms of evidence in ways that general models cannot achieve.

This approach represents a paradigm shift from the current practice in many academic fields that struggles to apply broad, shallow AI capabilities to specialized research domains. Instead of asking a general model to "understand" environmental policy through massive but unfocused training, our system will develop deep expertise in the specific discourses, power relationships, and knowledge systems that define environmental governance. The AI agent will recognize that community resistance language has evolved differently in different regions, that industry environmental impact assessments follow specific rhetorical formulas that vary by decade, and that federal agency justifications for environmental harm employ historically specific bureaucratic vocabularies.

A second dedicated Knight Lab studio of 18 senior students from majors across the university will work on AI agent development as their signature year-long Northwestern project, in a parallel team-based student capstone as the studio conducted for Phase 2. This studio will focus on developing the human-centered interfaces, visualization tools, and user interaction patterns that make sophisticated AI capabilities accessible to humanistic researchers, investigative journalism for ProPublica-type projects, and members of the public. James Lee will teach this Knight Lab studio in collaboration with Jeremy Gilbert, ensuring that students work directly with subject matter experts and PhD student researchers so that cutting-edge AI technology serves the specific needs of environmental humanities inquiry rather than generic information processing.

Additionally, a second Media, Publics and Rhetorics PhD student will be recruited with a dissertation focus on digital humanities and computational approaches to environmental archives. I will serve as dissertation advisor for this student, whose research will contribute to both technical development and scholarly analysis of how context engineering and MCP frameworks transform humanities research methodologies, creating new possibilities for domain-specific AI that serves scholarly inquiry rather than generic information processing.

The technical implementation team consisting of James Lee, Han Liu (Computer Science/Statistics), Aihan Liu (Data Scientist), and two PhD students will develop this context-aware system while simultaneously building MCP infrastructure that enables data flow between the dataset, public platform, and AI agent. This architectural innovation transforms static library collections into dynamic, interconnected data systems that support both historical research and computational analysis. The resulting infrastructure establishes a new model for "collections as data" that other institutions can adapt for their own specialized archives, representing a shift from broad AI applications toward deep, contextually sophisticated systems designed specifically for humanistic inquiry against library collections.

**Phase 4: Environmental Benchmarking Stream (Months 1-24)**

Recognizing the environmental implications of digital infrastructure, this stream continuously assesses and documents the carbon costs and sustainability decisions made throughout all other project phases. Kelsey Rydland (Director of Digital Scholarship) will lead a comprehensive monitoring framework using established environmental assessment tools to create real-time feedback loops that inform computational decision-making throughout the project. Environmental benchmarking of AI to date has focused on initial model training as a one time environmental impact, but little attention has been paid to the continuous, cumulative process of inference that expends power and resources every time a model is deployed for analysis. To our knowledge, this project would represent one of the first efforts to monitor and assess this incremental environmental cost of AI deployment over time.

The environmental monitoring approach will integrate multiple measurement tools to capture the full scope of the project's carbon footprint. **Codecarbon**, the widely-used open-source tool for tracking energy consumption during AI model training and inference, will monitor both the context engineering work in Phase 3 and the ongoing inference requests from public platform users. This tool automatically converts energy consumption to CO2 estimates, enabling continuous assessment of the AI agent's environmental impact as users interact with the archive.

For cloud infrastructure monitoring, the team will utilize **AWS Carbon Footprint Tool** to track the emissions associated with hosting, storage, and compute resources that support the OCR pipeline, dataset management, and AI/website operations. This cloud-based monitoring will provide institutional-level carbon accounting that complements the application-specific monitoring from Codecarbon.

**RAPL (Running Average Power Limit),** Intel's hardware-level processor energy monitoring interface, will enable precise measurement of energy consumption during local development and testing phases, particularly for the custom computer vision models that Aerith Netzer will develop for post-processing optimization. This hardware-level monitoring will provide granular data about the energy efficiency of different algorithmic approaches.

**Green Algorithms**, the web-based computational carbon footprint calculator, will serve as a benchmarking tool to validate our direct measurements and provide comparative analysis against standard computational approaches. This will help demonstrate how our specialized context engineering and MCP framework compares environmentally to traditional AI implementations.

This continuous monitoring approach will enable reflective decision-making about computational efficiency, server optimization, and sustainable practices rather than treating environmental impact as an afterthought or nuisance. The team will document carbon-aware computational practices, including strategies for reducing inference energy consumption, optimizing data processing workflows, and scheduling computationally intensive tasks during periods of renewable energy availability.

The deliverable from this phase will be a detailed case study documenting the environmental costs and benefits of large-scale digitization projects, with particular attention to the carbon implications of deep, specialized AI systems versus broad, general-purpose models. This documentation will serve as a model for environmentally responsible digital practice, ensuring the project's environmental benefits outweigh its computational footprint while establishing benchmarks that other institutions can use for their own environmental archive projects.

**Phase 5: Research Test Projects (Months 13-24)**

The final twelve months integrate all previous work streams through three interdisciplinary research initiatives led by faculty that test and refine the platform's capabilities while generating environmental research for academic impact and investigative journalism for public audiences. These projects will serve as proving grounds of sorts that stress test our assumptions in each of the previous four phases, while demonstrating the potential of our project to generate new research findings in published form.

The ***Community Voice and Resistance Analysis*** project, led by James Lee, will systematically extract and analyze the language of impacted communities, contained in letters, public comments, town hall transcripts, to surface recurring themes of resistance, adaptation, and environmental identity. By comparing these voices with government and industry rhetoric, this research will reveal patterns in how marginalized perspectives have shaped or been excluded from the environmental policy process, demonstrating the platform's capacity for large-scale rhetorical analysis and comparative study.

The ***Toxic Exposure at Military Bases*** investigation, led by Elizabeth Shogren, will enable veterans, scientists, and advocates to trace exposure to hazardous chemicals (TCE, PFAS) at federal sites, supporting medical research into unexplained illnesses, policy reform for cleanup standards, and personal redress through legal action. This project will test the platform's ability to connect textual evidence with geographic analysis and quantitative data extraction from technical documents, leveraging Shogren's extensive experience in environmental investigative journalism.

The ***Environmental Discourse Evolution*** study, led by Jeremy Gilbert, will compare EIS information about climate and ecosystems from 2004 to present against the rise of social media and internet discourse, examining how environmental knowledge and contestation have transformed across information ecosystems. This research will demonstrate the platform's temporal analysis capabilities and its potential for interdisciplinary scholarship connecting digital humanities with media studies and environmental communication.

**Integration Architecture and Timeline Coordination**

The MCP infrastructure enables continuous integration across all work streams, ensuring that improvements in data processing immediately benefit both public users and AI capabilities, while user feedback from the platform directly informs AI development priorities. This interconnected approach eliminates the traditional bottlenecks of sequential development, and reduces the risk of single points of failure disrupting the entire project, while enabling the specialization that the interdisciplinary scope of our project requires. The overlapping timeline structure ensures that each work stream can progress according to its own technical requirements while contributing to the overall system architecture that will establish new standards for environmental archive research and analysis.

**SUSTAINABILITY:**

This project serves as a strategic, signature initiative for Northwestern University Libraries, marking our first concerted effort to transform one of our most distinctive local collections into a dataset for broad public access and use. It is the cornerstone of a new library vision that reimagines collections as dynamic data, creating platforms for both computational analysis and public engagement. Northwestern University Libraries is deeply committed to the long-term preservation, accessibility, and future enrichment of this vital public resource. By establishing a new type of digital environmental archive, we aim for the platform to attract future research grants and partnerships that will ensure its continued growth and relevance. In brief, our project is not a time-bound undertaking but an investment by the library, and the University is committed to ensuring it continues to evolve and enhance public access for years to come.

**TECHNOLOGY:**

The proposed work involves the development and integration of three core technologies:

1. **Optimized OCR Processing:** We will move beyond standard digitization to utilize advanced Optical Character Recognition (OCR) technology specifically optimized for the complexities of historical government documents, which often include technical diagrams, handwritten notes, and oversized maps that are crucial for research but challenging for standard software.
2. **Public Exhibit Platform:** The public-facing website will be built on "Canopy," Northwestern's innovative, in-house platform for creating visually rich, narrative-driven digital exhibits. This technology has been selected because it moves beyond a simple database interface, allowing us to frame the archive as a site for storytelling and guided discovery, making it accessible to a broader public audience.
3. **AI Environmental Research Agent:** The most innovative technology is an AI-powered agent designed for humanistic inquiry. It will leverage natural language processing to enable intuitive, chat-based searching of millions of pages. It will employ text mining models to identify and compare the distinct rhetorical positions of government, industry, and community voices across the archive. Finally, it will use computer vision models to extract and structure quantitative data from thousands of non-textual sources like legacy paper maps and technical drawings, unlocking information that would otherwise be inaccessible. This technology was selected to allow researchers to ask new, large-scale questions that are impossible to answer through manual reading alone.

**POTENTIAL CHALLENGES:**

We anticipate several potential challenges and have designed our activities to mitigate them:

* **Technical Complexity:** The quality and format of historical documents vary greatly, which could pose challenges for automated OCR and data extraction. To address this, we are partnering with a vendor specializing in complex documents and have built in a rigorous quality control workflow led by our digitization experts. Similarly, developing an AI that provides nuanced, relevant results for humanistic inquiry is a significant challenge, which we are mitigating through a human-centered design process led by the Knight Lab and an iterative development cycle with user feedback.
* **Scale of Data:** Managing and processing a dataset of this size requires significant computational resources and robust data architecture. Our plan directly addresses this by allocating funds for a scalable cloud infrastructure and leveraging the expertise of our data scientists and curation specialists to design an efficient data management plan from the outset.
* **Ensuring Broad User Engagement:** A key challenge is ensuring the platform is truly useful for our diverse intended audience, from academic researchers to community activists. Our strategy addresses this by building the platform as an accessible "exhibit" rather than a technical database and by directly involving end-users in the design and testing process through pilot partnerships.

**CONCLUSION:**

In preserving and activating this archive, Northwestern University will illuminate the environmental conscience of U.S. environmental governance and provide essential tools for reckoning with its complex legacy. This project represents a unique opportunity to transform how we understand the intersection of governance, environment, and justice while creating infrastructure for environmental research and advocacy. With Mellon Foundation support, our initiative will help recover the oft-silenced voices that shaped America's environmental legacy, and empower new forms of public dialogue grounded in the historical record of resistance and debate.

APPENDIX FIGURES:

A screenshot of a computer

AI-generated content may be incorrect.