NEPATEC1.0: First Large-Scale Text Corpus of National Environmental Policy Act PDF Documents

Shivam Sharma, **Dan Nally,** **Mike J. Parker,** **Sai Munikoti,**

Sameera Horawalavithana*

Pacific Northwest National Laboratory, Richland, WA, USA

Abstract

An environmental impact statement (EIS) is a written document that contains detailed analysis of the potential environmental effects of a proposed major federal action. The preparation of an EIS and other procedural requirements of the National Policy Act (NEPA) are mainstays of federal decision-making and natural resource management. NEPA serves as a critical environment safeguard and opportunity for public engagement, while also facing scrutiny from efforts to streamline and expedite environmental permitting processes enabling the deployment of critical energy and infrastructure projects. Directed retrieval and interpretation of information contained in completed EISs, individually and in aggregate, could help improve the efficiency and outcomes of future NEPA reviews. To encourage developers to build AI tools with this objective, we release a text corpus of NEPA PDF documents, **N**ational **E**nvironmental **P**olicy **A**ct **Te**xt **C**orpus (NEPATEC1.0). NEPATEC1.0 consists of textual data extracted from more than 28k EIS documents associated with 2,917 projects reviewed under NEPA. This textual data consists of page-wise content from each of the documents and a set of named entities flagged from the page-wise text. In addition, we organize the documents by the level of projects and enrich with metadata (e.g., project title, agency, and location).

1 Introduction

The National Environmental Policy Act of 1969, as amended (NEPA), is a bedrock and enduring environmental law in the United States with the express intent of fostering a productive harmony between humans and the environment for present and future generations. The NEPA statute (42 U.S. Code 4321 et seq.) and implementing regulations of the Council on Environmental Quality (40 Code of Federal Regulations parts 1500 through 1508) establish procedures requiring all federal agencies to consider environmental effects in their planning and decisions and to inform the public. As a first step, federal agencies must determine whether NEPA applies to a proposed action and then determine the appropriate level of environmental review. A categorical exclusion is the most basic level of NEPA review and addresses those categories of actions that do not individually or cumulatively have a significant effect on the environment. An environmental impact statement (EIS) is the most detailed level of NEPA review and is required for major federal actions with significant environmental effects. If it is unknown whether a proposed action has the potential to have a significant effect on the environment, an agency must first prepare a more concise document called an environmental assessment (EA) to support its determination (Figure 1).

*Please contact policyai@pnnl.gov, and visit `https://www.pnnl.gov/projects/policyai` for more details.

Preprint. Under review.

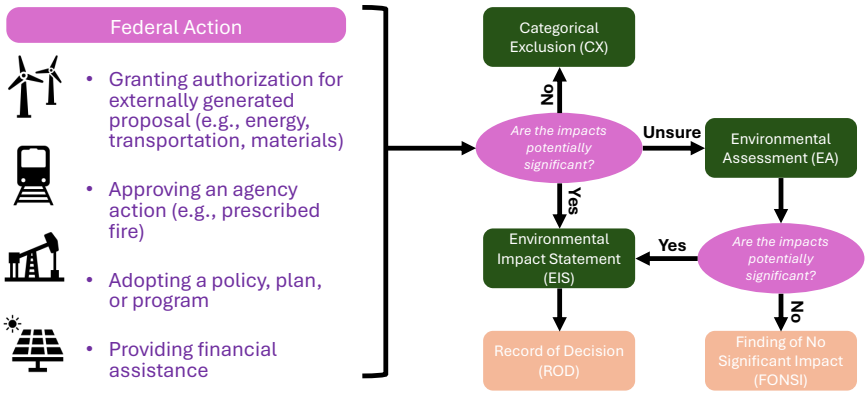

Figure 1: NEPA Decision Making Process and Lifecycle

Each type of NEPA review requires preparation of a written document disclosing relevant information that supports the agency's decision-making process. Recent changes to NEPA now limit EAs to 75 pages and EISs to 150 pages, excluding citations, appendices, and information displayed graphically. Historically, most EISs have been substantially longer than 150 pages. Average document length for EISs sampled by the Council on Environmental Quality from 2013 to 2018 was 575 pages for draft documents and 661 pages for final documents (excluding appendices, which accounted for, on average, another 584 pages and 1,042 pages, respectively) [3]. An agency typically begins the NEPA process after determining the appropriate level of NEPA and establishing that there is an adequate amount of information available about the proposed action and that any other applicable application requirements are met, which may require several iterations. Each EIS must be released as a draft for public comment and then as a final document, with at least two opportunities for persons and organizations that may be interested in, or affected by, the proposed action to provide comment (Figure 2). Less commonly, in event a proposed action changes significantly or circumstances change considerably, agencies may prepare a supplement to a draft or final EIS. At the conclusion of a NEPA process requiring an EIS, an agency publishes a record of decision stating the decision, identifying the alternatives analyzed, and any required mitigation. Each EIS is orbited by a host of separate but related satellite documents and multimedia files, such as references cited, copies of other permits and authorizations, baseline data and project-specific data analysis, geospatial data, Federal Register notices, and public outreach materials.

Although the majority of proposed actions reviewed by agencies are addressed through categorical exclusions, major federal actions that require an EIS are the most conspicuous and information-rich products of the NEPA process. Major federal actions may include granting an authorization for an externally generated proposal, approving an agency action, providing financial assistance, or adopting a policy, plan, or program. Common examples of major federal actions include approving permit and right-of-way applications for energy development projects or construction of roads or transmission lines across public lands.

All EISs share common elements required by NEPA, but there is no universally standardized format for organizing the information contained in an EIS. Although a higher degree of standardization may be desirable for those preparing and reviewing NEPA documents, the diversity of proposed actions and unique regulatory framework require flexibility to accommodate. Comparing the tables of contents for EISs prepared by different agencies at different times, specific to different locations and types of actions exhibits a wide variety of approaches. As long-format documents, the content of each EIS (including appendices) may be spread across multiple volumes. Although EISs consist predominately of textual data, they also contain information in figures, tables, and cited references.

Previous works have identified the challenges in cataloging larger EIS collections [2, 1]. They primarily focused on improving the metadata used to annotate the documents such as identifying lead agencies, document version, date, agency, and states [2]. In addition, Bethard et al. [1] evaluate the performance of rule-based baselines in aligning multiple document versions to the project, and detecting reused text in between the draft and final EIS versions.

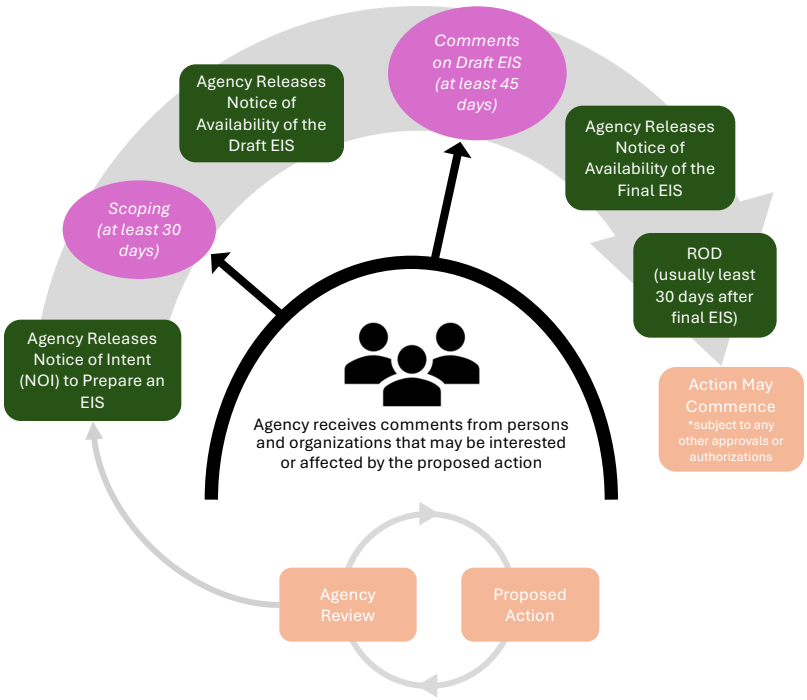

Figure 2: EIS Preparation process

In this work, we publicly released a large text-corpus of NEPA EIS documents, named as **N**ational **E**nvironmental **P**olicy **A**ct **Te**xt **C**orpus (NEPATEC1.0)[2]. NEPATEC1.0 contains textual data from draft, final, and supplemental EIS documents organized by the project. We describe the data collection and processing steps (Section 2) and the potential technical challenges that can be addressed using the dataset (Section 3).

2 NEPATEC1.0 Construction, Processing and Augmentation

NEPATEC1.0 is an AI-ready dataset that contains data extracted from the Environmental Impact Statement (EIS) Database provided by U.S. Environmental Protection Agency (EPA). An EIS is a government document that analyzes the potential environmental effects of a proposed project and identifies ways to mitigate those effects. NEPATEC1.0 contains textual data and associated metadata extracted from 2,917 projects. These projects contain 28,212 documents, 4.6 million pages, and over 3.6 billion tokens of textual data (using GPT2 tokenizer). In this section, we describe the dataset collection and preprocessing steps (Section 2.1) and the metadata mapping process (Section 2.2) as outlined in Figure 3.

2.1 Dataset Collection and Preprocessing

We outline the data processing and augmentation pipeline in Figure 3. The NEPATEC 1.0 dataset was scraped from the EPA data website by making an empty search[3], which returned all the PDF links in the database along with the document titles. In the course of our analysis, we discovered that some EIS document titles were duplicates, each linked to a different set of documents. We attempted to resolve this issue by merging the documents based on these duplicate titles. This process reduced the original collection from 12k EIS project document sets to approximately 7k. Further analysis revealed the existence of EIS projects with titles that were not exact duplicates but partially similar. To address this, we employed fuzzy matching techniques, which further reduced the 7k document sets down to 5k distinct sets of EIS PDFs.

[2] https://huggingface.co/datasets/PolicyAI/NEPATEC1.0

[3] https://cdxapps.epa.gov/cdx-enepa-II/public/action/eis/search

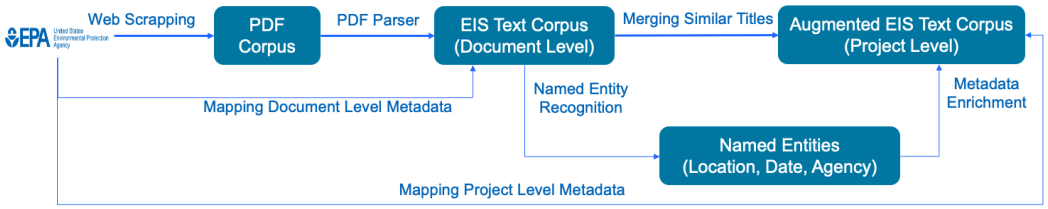

Figure 3: Data Processing and Augmentation Pipeline

We used PyMuPDF to parse textual and image data from the downloaded PDFs. The parsed textual data was split by pages. During analysis of these parsed outputs, we noticed around 2k PDFs either did not contain any textual information, or the parser was unable to parse the text in those PDFs. We used this page-wise text to extract named entities using the GLiNER toolkit [4]. The GLiNER model accepts around 400 tokens per query; thus, we processed and split the page-wise text to 150 words per batch and passed these batches through the GLiNER pipeline. The GLiNER pipeline accepts user-defined entities and can flag a word or set of words corresponding to an entity along with a confidence score representing the probability of the flagged word or set of words belonging to the labeled entity. In this work, we extracted five different entities from the page-wise text passed through the document parser. They are *Name (e.g., people or project), Date, Agency, Title,* and *Location (i.e., street, city, county, state, and country).*

2.2 Metadata Enrichment

After parsing the textual information and extracting the named entities from the textual content of the downloaded documents, we enriched the dataset with document-level metadata from the EPA website. The EPA website provides a list of five different metadata for each project and corresponding document set, namely:

1. **Agency:** List of agencies associated with the project
2. **State:** List of states
3. **Document Version:** Version of the document, whether the document set belong to the draft, final, or any other version of the EIS project stage
4. **EPA Comment Letter Dates:** List of EPA Comment Letter Dates
5. **Federal Register Date:** List of Federal Register Date

We faced similar issues as before in this section of the dataset creation, where there were partial duplicate titles between the downloaded document set and the metadata files. To mitigate this issue, we performed exact and fuzzy matching approaches, as discussed above, to map the metadata to their corresponding document sets. However, even after fuzzy matching, from a set of 5k downloaded set of EIS project documents, we were able to map only nearly 4k document sets to their corresponding metadata. Hence, from the total of over 35k downloaded PDFs, we were able to map only 30k PDFs to their corresponding project-level metadata. By project-level metadata, we mean to imply that we have no exact way to map document-level metadata, like dates or document versions, to their corresponding documents, hence, in NEPATEC 1.0, we dropped the document level metadata and kept all the dates in a list as project-level metadata.

Table 1 shows the basic statistics of the dataset. While the majority of the projects are from the Forest Service, the U.S. Army Corps of Engineers has the largest collection of EIS documents by token count for a single agency.

3 Technical Problems

Use of LLMs offers myriad opportunities to gain valuable insights through reading, interpretation, and analysis of the NEPA documents in NEPATEC1.0, as well as the ability to inform specialized generative tasks to support environmental permitting processes. The following list of opportunities highlights areas of particular interest to our research team and domain experts but is not exhaustive.

Table 1: NEPATEC 1.0 Statistics by Agency. This table showcases the top 20 agencies, sorted by number of projects, and their corresponding statistics, along with the rest of the agencies being grouped into the "Other" category. Because some projects overlap between agencies, the sum of these statistics do not represent the original split.

Agency Name	#Projects	#Documents	#Pages	#Tokens (In Millions)
Forest Service	682	3,403	434,279	337
Bureau of Land Management	323	3,620	464,198	353
U.S. Army Corps of Engineers	319	4,849	836,108	736
Federal Highway Administration	283	4,423	601,127	483
Federal Energy Regulatory Commission	145	892	176,634	119
National Park Service	132	286	63,890	47
National Oceanic and Atmospheric Administration	110	383	98,460	101
Fish and Wildlife Service	93	595	116,161	80
Bureau of Reclamation	75	860	265,819	249
Department of Energy	69	614	123,477	108
Nuclear Regulatory Commission	62	161	49,493	38
Federal Transit Administration	62	1,372	302,166	279
United States Navy	53	241	116,528	103
United States Air Force	47	297	85,387	55
Bureau of Indian Affairs	44	456	91,148	81
National Marine Fisheries Service	42	273	142,900	116
Federal Railroad Administration	37	1,210	88,007	65
Bureau of Ocean Energy Management	36	220	78,664	70
United States Army	35	168	24,871	15
Tennessee Valley Authority	34	105	29,762	20
Other	396	5,875	945,412	840

Extracting multi-level metadata. The length, complexity, and variable format of EIS documents presents a barrier to performing robust statistical analysis on the NEPATEC1.0 documents without further processing. However, AI-driven data processing could automate initial harvesting of a standardized set of metadata from all documents for human review and confirmation, saving considerable time over a fully manual process. Examples of metadata that may be useful for analysis include the type of proposed action (e.g., construction of a new liquid natural gas pipeline, adoption of a resource management plan, renewing a nuclear power plant operating license, or an agency rulemaking) and dates of major project milestones (e.g., notice of intent, publication of draft and final EIS, record of decision). Ultimately, metadata may be grouped as a series of related objects, such as metadata about the action under review (i.e., the type of proposed action), process (i.e., level of NEPA review), environmental resource area (e.g., air quality or terrestrial ecology), and document (i.e., type and version).

Extracting detailed project location data. Detailed geospatial information about a project's location is a specialized type of metadata that unlocks a wide array of search, analysis, and localization capabilities. For projects with defined footprints or activity areas (e.g., such as a transmission line maintenance right-of-way or recreation management area), AI-driven data processing could assist with the delineation of precise location information from textual and graphic data contained in the EISs, such as narrative descriptions of the project locations and maps. Such data could be stored in a file format with the capability to be displayed in mapping applications and analyzed using geospatial analysis software, such as ESRI Geodatabase, GeoJSON, or KML.

Metrics to evaluate NEPA efficiency Calls to streamline and improve the efficiency of NEPA and other permitting processes have been heard from industry and agency leaders, lawmakers, and the public. Although various laws, regulations, and initiatives have been launched for this purpose, there is limited information available to assess the actual effects on the permitting process. A few of many potential trends of interest include the length of EISs and appendices over time, length of time needed to complete the NEPA process, and trends in types of projects subject to NEPA review over time. Previously, studies of this type have required manual analysis of a sample of EISs [3]. Time needed to complete NEPA and other permitting processes are tracked for certain Federal infrastructure projects

on the Permitting Dashboard[4] maintained by the Federal Permitting Improvement Steering Council; however, a large number of NEPA reviews are not tracked in this manner.

Spatio-temporal trend analysis. Existing metadata or an expanded set of metadata obtained from the actions described above provides a unique dataset for spatio-temporal trend analysis. A centralized EIS database offers an opportunity to elucidate trends within or between specific geographic areas. For example, in what states or counties are solar energy projects under NEPA review? Or, chart how many oil and gas development projects triggered NEPA review (and indicate which levels of review) in the state of Wyoming over the last two decades.

Multi-document concept summarization. There is considerable variability in the format and content from one EIS to another, including common elements such as the description of the proposed action, alternatives, types of resources analyzed, analysis approach, mitigation requirements. Locating, reading, and interpreting a large number of EISs to compare and contrast different approaches is time consuming and difficult to perform in a systematic manner. Enlisting AI tools may enable NEPA document preparers to survey a larger number of documents and review a concise summary of differences that may enhance their understanding existing documents and inform their approach to drafting new documents. Example use cases include asking AI to provided an annotated list of the primary types of alternatives considered for wind energy projects in the Atlantic Ocean, or generating a list of potential mitigation measures for energy projects that may affect greater sage-grouse.

Language modeling. CEQ instructions require that EISs be written using plain language and "clear prose" (40 CFR 1502.8). LLMs have the capability to learn and mimic the style of EISs, and to edit existing text or generate new draft text to achieve a more consistent and simple writing style to improve reader experience and comprehension.

Visualization. The long and predominantly textual format of EISs makes them time consuming to read and digest, and nearly impossible to interact with on cell phones and other small mobile devices. CEQ regulations also encourage the use of visual aids or charts to help improve understanding (40 CFR 1508.8). Developing AI tools to visualize specific subsets of information graphically could aid in communicating content more instantaneously and in a more compelling and shareable format. This may allow information to reach a broader demographic and serve as a gateway for those willing to explore the written products in more detail. Potential examples range from a basic visualization on the frequency of different issues being raised in public comments to complex numerical taxonomy clustering.

Identifying scientific concepts (climate change, greenhouse gas emissions). Scientific studies play a crucial role in providing baseline information about the current state of the environment and in assessing potential project impacts. Depending upon the project type, location, and potential environmental effects, and with consideration for changing global environmental conditions, various types of scientific studies are cited in NEPA documents. It would be interesting to explore what scientific concepts and specific studies are being cited, how they change over time, and their variation across different dimensions such as project type, agency, and location. Such an exploration can help in identifying the right scientific studies for future reviews and highlight any gaps in existing research that need more attention for accurately assessing environmental impacts.

4 Acknowledgements

This work was supported by the Office of Policy, U.S. Department of Energy, and Pacific Northwest National Laboratory, which is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract DE-AC05–76RLO1830. This technical report has been cleared by PNNL for public release as PNNL-36124.

References

[1] S. Bethard, E. Laparra, S. Wang, Y. Zhao, R. Al-Ghezi, A. Lien, and L. López-Hoffman. Inferring missing metadata from environmental policy texts. In *Proceedings of the 3rd joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 2019.

[4] `https://www.permits.performance.gov/`

[2] E. Laparra, A. Binford-Walsh, K. Emerson, M. L. Miller, L. López-Hoffman, F. Currim, and S. Bethard. Addressing structural hurdles for metadata extraction from environmental impact statements. *Journal of the Association for Information Science and Technology*, 74(9):1124–1139, 2023.

[3] C. on Environmental Quality (CEQ) within the Executive Office of the President. `https://ceq.doe.gov/docs/nepa-practice/CEQ_EIS_Length_Report_2020-6-12.pdf`, 2024. [Online; accessed 21-June-2021].

[4] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois. Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023.