NEPA Project Analysis: Clean Energy Environmental Reviews

National Environmental Policy Act Text Corpus (NEPATEC) 2.0 Analysis

Your Name

2026-01-16

Table of contents

1 Project Overview

This document is a good reference for how to contextualize this project, the project goals and data, and to clearly understand the project deliverables and timeline.

1.1 Project context: National Environmental Policy Act (NEPA)

Federal environmental permitting—and, relatedly, the National Environmental Policy Act (NEPA)—has often been blamed as a core contributor to delays in infrastructure deployment. However, research to date on NEPA has been hindered by federal agency data management practices: NEPA documents are scattered across numerous agency databases, often in machine-unreadable formats, and typically lack basic metadata and other identifiers. As such, researchers have had to craft bespoke subsets of NEPA data from which to glean insights—a time-consuming task that has limited the information available about NEPA's effectiveness and paved the way for a national permitting reform conversation rife with anecdotes and cherry-picked information.

The newly released National Environmental Policy Act Text Corpus (NEPATEC) 2.0 dataset from the Pacific Northwest National Laboratory's (PNNL's) PermitAI project has the potential to add new, more comprehensive evidence into this conversation. PNNL has built and released a comprehensive dataset of past environmental reviews and permitting documents containing millions of pages and billions of words in machine-readable JSON format. The database contains more than 120,000 NEPA documents[1] from 60,000 projects prepared by more than 60 different agencies. Each document contains metadata for (as applicable):

- Lead agency
- Category
- Type of review
- Name of project
- Location
- Project sponsor
- Project sector
- Project type (a subset of project sector)
- Type of document
- Document title
- Agency or contractor responsible for preparing the document
- Categorical exclusion category
- Summary of the proposed action

[1] categorical exclusions, draft and final environmental assessments, draft and final environmental impact statements, records of decision, findings of no significant impact, and other supporting documentation

> **Note**
>
> The National Environmental Policy Act Text Corpus (NEPATEC) 2.0 can be accessed here on huggingface.

1.2 Data structure

This shows the original data structure of the project in its original json format. The code in this repo manipulates this raw data.

```
{
    "project": {
        "project_ID": "UNIQUE PROJECT ID FOR PUBLIC VERSION",
        "project_title": {
            "value": ""
        },
        "project_sector": {
            "value": ""
        },
        "project_type": {
            "value": ""
        },
        "project_description": {
            "value": ""
        },
        "project_sponsor": {
            "value": ""
        },
        "location": {
            "value": ""
        }
    },
    "process": {
        "process_family": {
            "value": ""
        },
        "process_type": {
            "value": ""
        },
        "lead_agency": {
            "value": ""
```

```
        }
    },
    "documents": [
        {
            "metadata": {
                "document_metadata":{
                    "document_ID": {
                        "value": "UNIQUE DOC/FILE ID FOR PUBLIC VERSION"
                    },
                    "document_type": {
                        "value": ""
                    },
                    "document_title": {
                        "value": ""
                    },
                    "prepared_by": {
                        "value": ""
                    },
                    "ce_category": {
                        "value": ""
                    }
                },
                "file_metadata":{
                    "file_ID": {
                        "value": "UNIQUE DOC/FILE ID FOR PUBLIC VERSION"
                    },
                    "file_name": {
                        "value": "PDF NAME"
                    },
                    "section_or_volume_title": {
                        "value": ""
                    },
                    "main_document": {
                        "value": ""
                    },
                    "total_pages": {
                        "value": ""
                    },
                    "file_provider": {
                        "value": ""
                    }
```

```
                }
            },
            "pages": [
                {
                    "page number": 1,
                    "page text": "PAGE 1 TEXT"
                },
                {
                    "page number": 2,
                    "page text": "PAGE 2 TEXT"
                }
            ]
        }
    ]
}
```


2 Project deliverables

For this project, we want to create tables, figures, and maps that help us learn about the data and answer the following questions:

2.1 Phase 1

Table 1: Phase 1 Deliverable Timeline

Deliverable	Due_Date
1. Clean Energy Projects	Jan 23, 2026
2. Programmatic Reviews	Feb 6, 2026
3. CE vs EA vs EIS	Jan 23, 2026
4. Geography	Feb 6, 2026
5. Pages Over Time	Feb 20, 2026
6. Technology-Specific	Feb 20, 2026

1. **Data on number of clean energy projects within the dataset**: number of projects broken down by technology (e.g., offshore and onshore wind, solar, geothermal, nuclear), lead agency, and location
2. **Data on programmatic and tiered reviews**: how many tiered reviews are there compared to total and are they completed faster

3. **Data on how many clean energy projects have been categorically excluded vs. have required environmental assessments and environmental impact statements**
    - Broken out by number of projects, generation capacity, and change over time
4. **Data on geography/project location**: whether projects are multi-state or multi-agency
5. **Number of pages over time**, including pre- and post- Fiscal Responsibility Act of 2023 (FRA), which set page limit requirements
6. **Technology-specific inquiries**
    - Transmission lines: length of lines from project summary correlated with timelines, location, etc.
    - Geothermal: timelines of environmental reviews for different phases of the same project
    - Carbon and hydrogen pipelines (if available): length of pipelines from project summary correlated with timelines, location, etc.; compare to natural gas pipelines

2.2 Phase 2

Phase 2 of the project is still not settled and we'll need to scope it based on what we find and do in Phase 1. Nevertheless, these are the envisioned deliverables:

1. **Reasons why NEPA was triggered** (e.g., federal land, federal funding) for different types of projects
2. **Determinations of significance across resource areas**; factors that contribute to a determination of "significant impact"
    - Starting with mitigated FONSIs
3. **Differences and similarities between NEPA reviews** for fossil fuel and clean energy projects, as well as linear and non-linear projects—application of categorical exclusions, timelines, geography, etc.
4. **Timelines for categorical exclusions, environmental assessments, and environmental impact statements**, including segmentation by years (e.g., pre- and post-FRA [which set timelines for reviews], different CEQ NEPA regulations, agency, and type of project)
    - Timeline outliers could then be investigated through a case study approach to identify contributing factors, including whether NEPA was a cause of delay or not

- May need to cross-reference with the Notice of Intents in the Federal Register using their API to get the start date

5. **Technical support for new regulatory categorical exclusion development**: identifying patterns in FONSIs

3 Project deliverable timelines

Table 2: Project Timeline and Deliverables

Meeting	Date	Deliverables
Kickoff	Jan 9, 2026	(0) Build database
1	Jan 23, 2026	(1) Clean energy projects(3) CE vs EA vs EIS
2	Feb 6, 2026	(2) Reviews(4) Geography
3	Feb 20, 2026	(5) Pages(6) Technology
4	Mar 6, 2026	Present all findings

4 Detailed Deliverable Specifications

4.1 Deliverable #1: Clean Energy Project Analysis

I envision making three tables to satisfy this deliverable. All of the variables for all the tables of deliverable 1 are in the metadata.

> ⚠ Warning
>
> **Data Challenge**: Many projects have multiple tags. For example, the project named as "Coldfoot Cell Tower" in the `project_title` variable has two values for `project_type`: ['Utilities (electricity, gas, telecommunications)' 'Broadband']. While "Utilities (electricity, gas, telecommunications)" is a Clean Energy category, "Broadband" is not.

Solution approach: Create a new variable flag, called `energy_type_questions`. Any clean energy tag that also has a fossil fuel tag should be classified as fossil fuel.

Required new variables: - `project_type_count`: count of unique values for `project_type`
- `project_type_count_clean`: counts unique clean energy tags
- `project_type_count_fossil`: counts unique fossil fuel tags

4.1.1 Clean Energy by Technology

```
# Placeholder for clean energy by technology analysis
# This table will group the three datasets together (ER, ES, and CE) and create counts by
```

Required variables: - `project_type` - `project_energy_type`

Technology	Environmental Review	Environmental Statement	Categorical Exclusion
Offshore wind	XXX	XXX	XXX
Onshore wind	XXX	XXX	XXX
Solar	XXX	XXX	XXX
Geothermal	XXX	XXX	XXX
etc.	XXX	XXX	XXX
Total	**XX,XXX**	**XX,XXX**	**XX,XXX**

4.1.2 Clean Energy by Lead Agency

```
# Placeholder for clean energy by lead agency analysis
```

Required variables: - `lead_agency` - `project_energy_type`

Agency	Environmental Review	Environmental Statement	Categorical Exclusion
Department of the Interior	XXX	XXX	XXX
Department of Energy	XXX	XXX	XXX
Major Independent Agencies	XXX	XXX	XXX
Department of Agriculture	XXX	XXX	XXX
etc.	XXX	XXX	XXX
Total	**XX,XXX**	**XX,XXX**	**XX,XXX**

4.1.3 Clean Energy by Location

```
# Placeholder for clean energy by location analysis
# Will need to create project_state and project_county variables from project_location
```

> **Note**
>
> **Data processing needed**: From `project_location`, we'll need to create `project_state` and `project_county` variables. The values for `project_location` are not clean or consistent and will need extra work to create county or state-level classifications.

Required variables: - `project_location` - `project_state` - `project_county`

State	Environmental Review	Environmental Statement	Categorical Exclusion
Alaska	XXX	XXX	XXX
Arizona	XXX	XXX	XXX
California	XXX	XXX	XXX
Colorado	XXX	XXX	XXX
etc.	XXX	XXX	XXX
Total	**XX,XXX**	**XX,XXX**	**XX,XXX**

4.2 Deliverable #3: Project Status Analysis

I envision making three tables:

4.2.1 Project Status by Energy Type

```
# Create project_energy_type from project_type variable
# Match values against Clean and Fossil Energy tags in ea_project_type.txt
```

Required variables: - `project_type` - `project_energy_type` - `project_energy_type_questions` - `project_type_count` - `project_type_count_clean` - `project_type_count_fossil`

Energy Type	Environmental Review	Environmental Statement	Categorical Exclusion
Clean Energy	XXX	XXX	XXX
Fossil Fuel	XXX	XXX	XXX
Other	XXX	XXX	XXX
Total	**XX,XXX**	**XX,XXX**	**XX,XXX**

4.2.2 Clean Energy Projects by Project Status and Generation Capacity

```
# Extract generation capacity from document text
# Two potential approaches:
# 1. Specify possible ways projects mention generation capacity and use regex/pattern matc
# 2. Use LLM to read text documents and extract generation capacity information
```

> **!** Important
>
> **Technical Challenge**: Generation capacity needs to be extracted from document text. Two approaches: 1. **Pattern matching**: Specify possible ways projects mention generation capacity and use Python sets to search documents 2. **LLM approach**: Feed text documents into an LLM to extract generation capacity information
> Each page is in text format in the `page_text` variable, grouped by `project_id`.

Required variables: - `project_id` - `page_text` - `project_gencap_measure` (unit: MW, kW, etc.) - `project_gencap_value` (numeric value)

Energy Generation Capacity	Environmental Review	Environmental Statement	Categorical Exclusion
High	XXX	XXX	XXX
Medium	XXX	XXX	XXX
Low	XXX	XXX	XXX
Total	**XX,XXX**	**XX,XXX**	**XX,XXX**

4.2.3 Clean Energy Projects by Project Status and Change Over Time

```
# Complex timeline construction from documents
# Projects may move from CE to ER to ES (more intensive review)
# Need to check which projects are in multiple datasets
```

> **⚠** Warning
>
> **Most Complex Variable**: This requires reading all documents associated with a project to construct a timeline as it moves through the NEPA process. Projects in CE won't be in other datasets, but projects in ER could move to ES.
> **Questions for implementation**: - How to organize timeline variable (list, vector, etc.)? - Create multiple views: by year, by review duration, by number of days under review

Required variables: - `project_id` - `page_text` - `timeline` (structure TBD - need suggestions)

Year	Environmental Review	Environmental Statement	Categorical Exclusion
2025	XXX	XXX	XXX
2024	XXX	XXX	XXX
2023	XXX	XXX	XXX
Total	**XX,XXX**	**XX,XXX**	**XX,XXX**

5 Project-level rules and procedures

5.1 Variable naming

Ensure where possible that the name of any newly created variables include a prefix in the variable name indicating the level they are. All variables at the "document" level begin with `document_[varname]`, "project" level begin with `project_[varname]`, etc.

Universe of possible prefixes: - `project_` - project level variables - `process_` - process level variables
- `document_` - document level variables - `pages_` - page level variables

> **Note**
>
> Most new variables will likely be at the "project" stage since that is the level of observation for most analyses in this project.

6 Code Structure

```
# Placeholder for main analysis code structure
# Load data
# Clean and process variables
# Create derived variables
# Generate tables and visualizations
# Export results
```