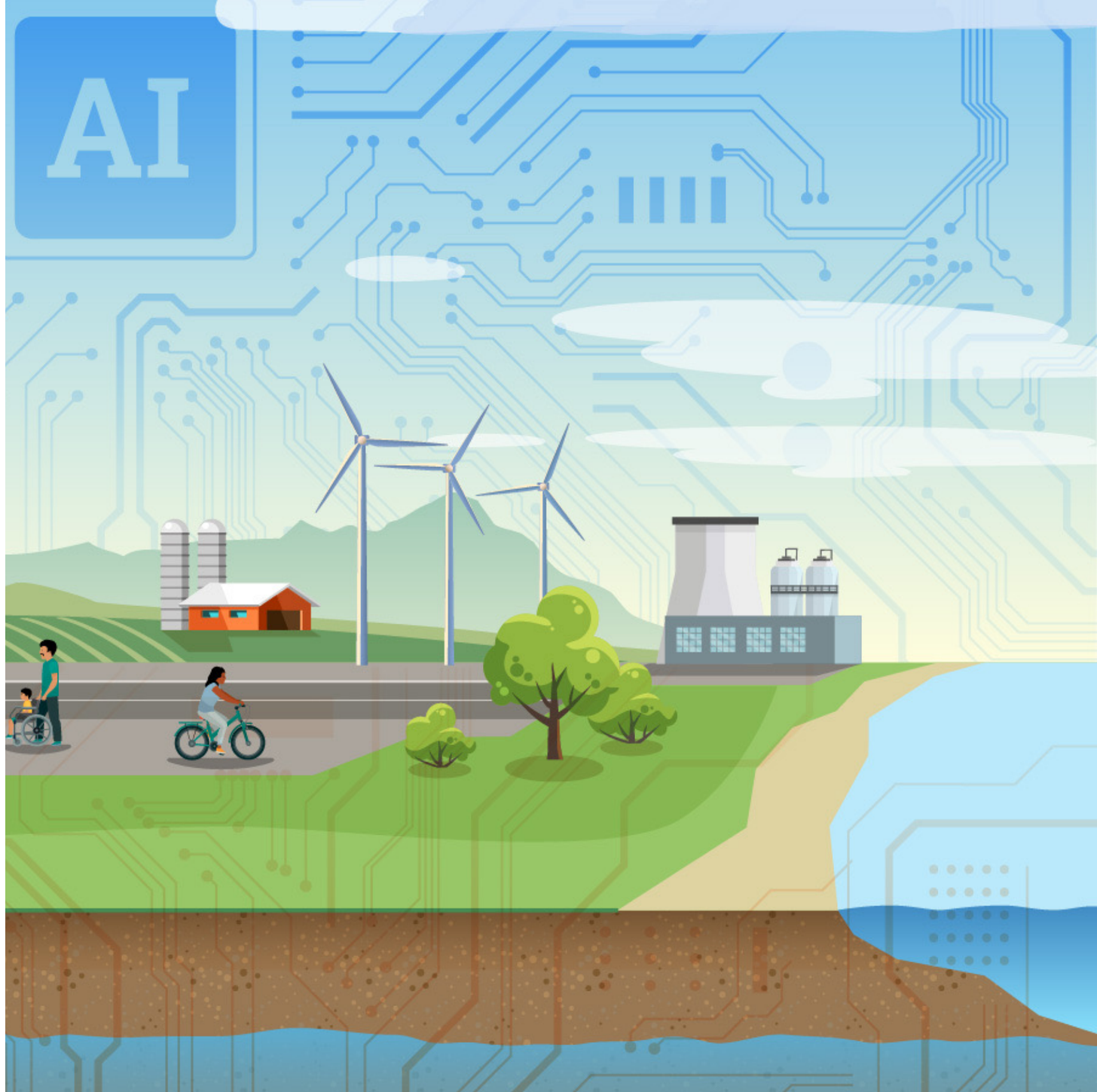
Pacific Northwest
NATIONAL LABORATORY

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

Dan Nally
NEPA Project Manager

Shivam Sharma
Data Scientist



U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

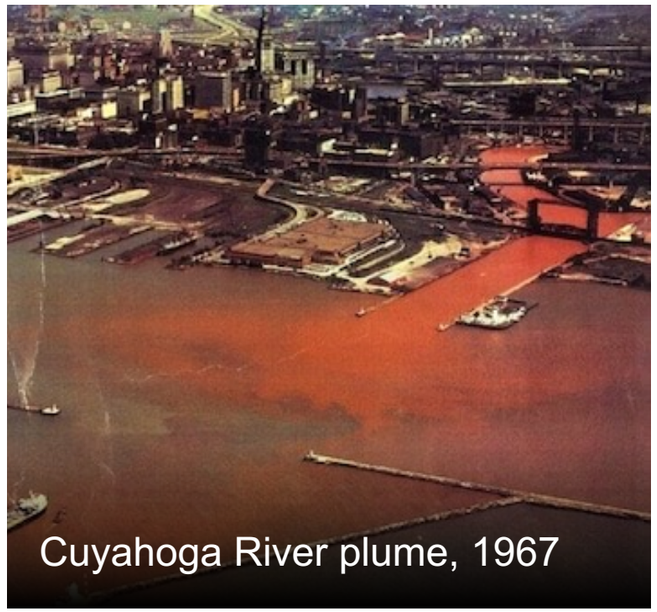
PNNL-SA-200118

Origin and Purpose of NEPA

- Enacted in 1969; signed into law in 1970
- Landmark and often replicated environmental law
- Requires Federal agencies to consider significant environmental consequences of their proposed actions and inform the public
- Plays critical role in promoting sound decision making and reducing damage to the environment

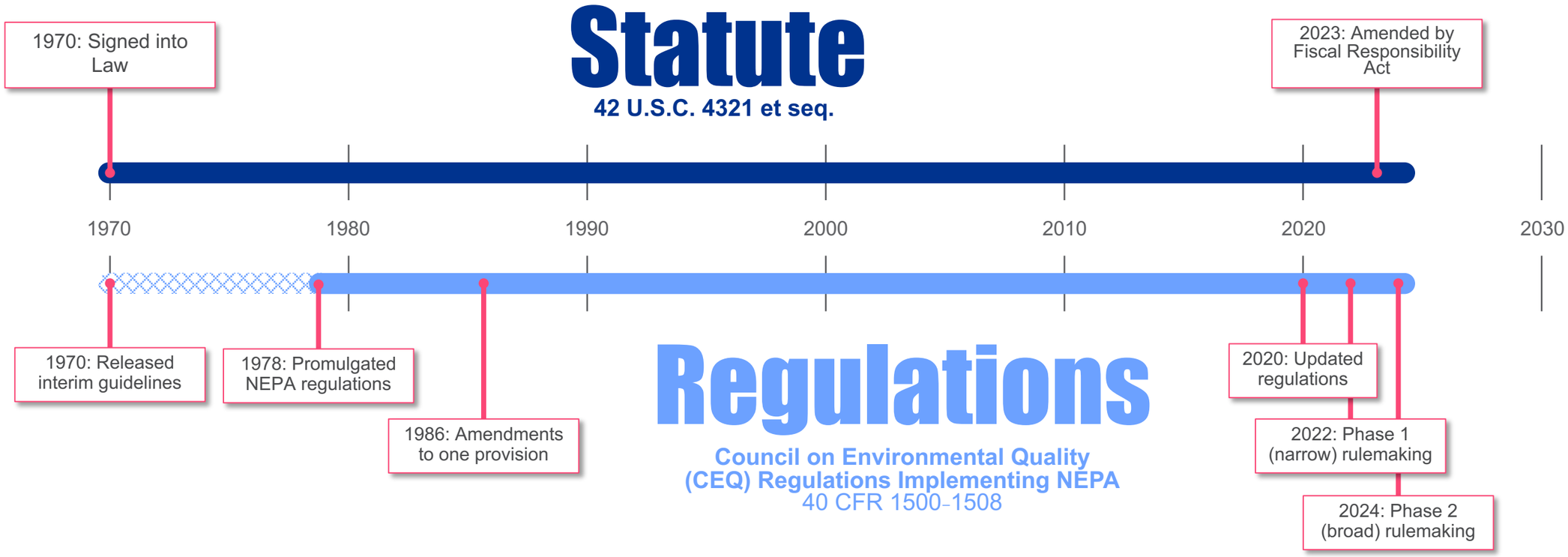

Apollo 11, 1969


President Nixon, 1970


Cuyahoga River plume, 1967


Santa Barbara oil spill, 1969

Substantial Changes to NEPA



Note: Minor technical and typographical changes not shown.

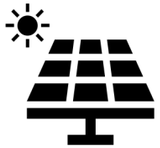
The NEPA Process

Federal Action

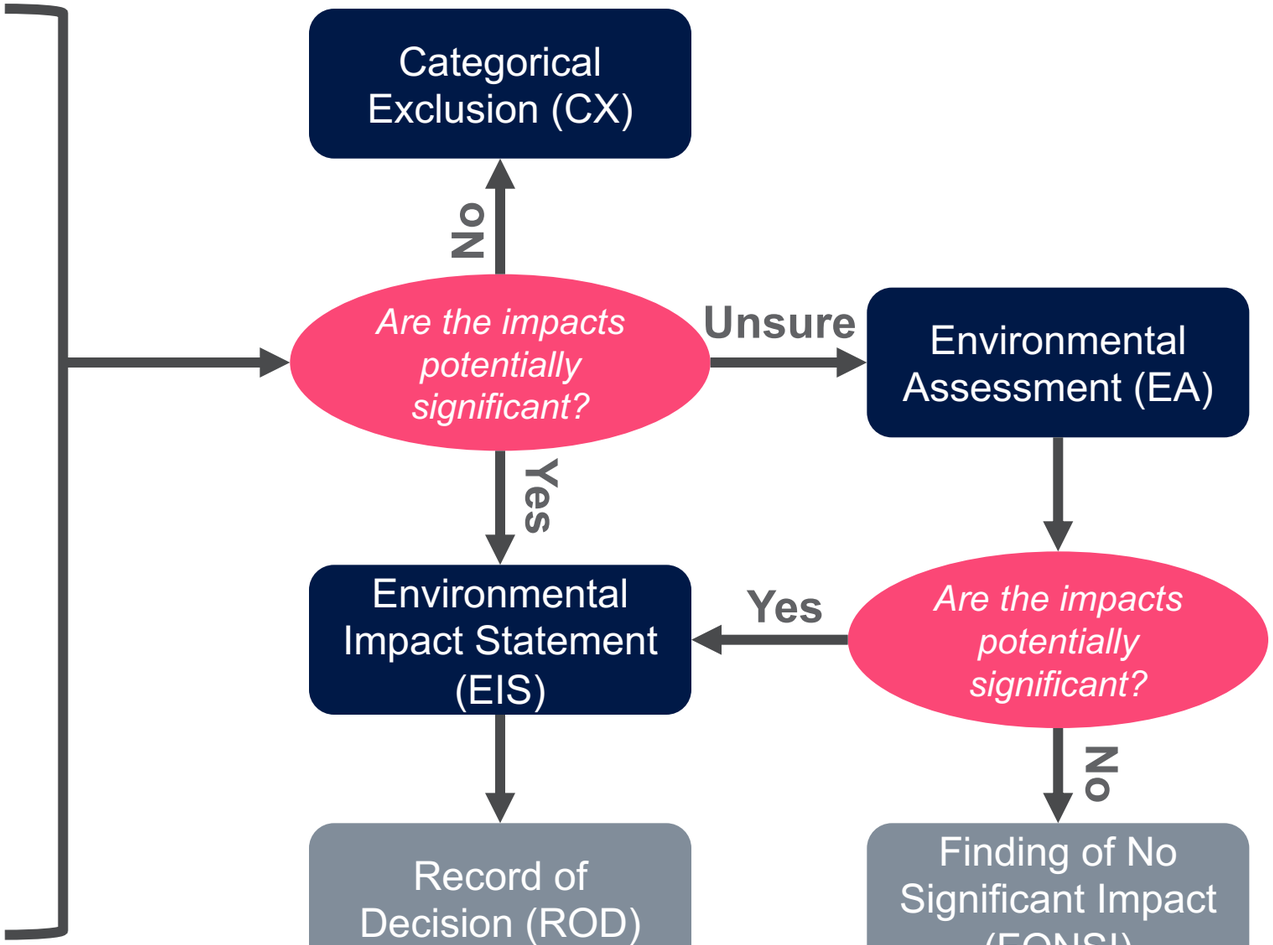
- Granting authorization for externally generated proposal (e.g., energy, transportation, materials)

- Approving an agency action (e.g., prescribed fire)

- Adopting a policy, plan, or program

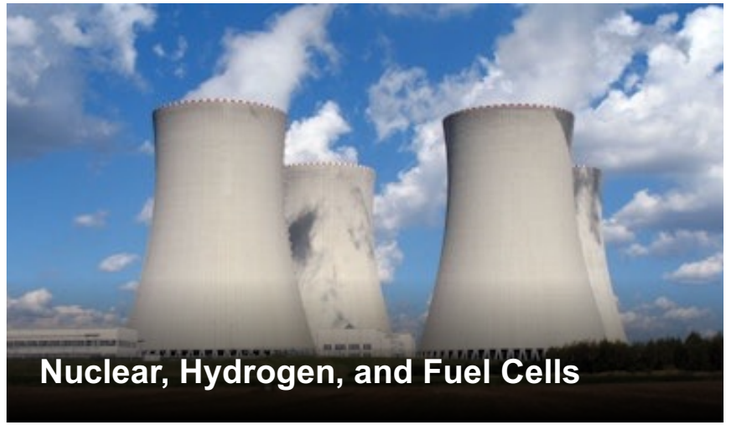


- Providing financial assistance



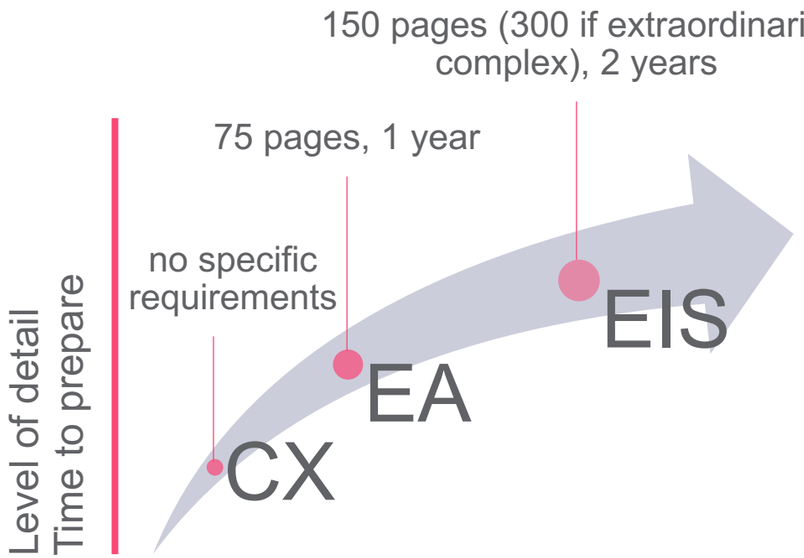
NEPA for Clean Energy Actions

- Commercial-scale projects often involve Federal lands or financial assistance, triggering NEPA requirements



Timelines by Document Type

Regulatory Requirements



Source of regulatory requirements: 40 CFR 1501.10

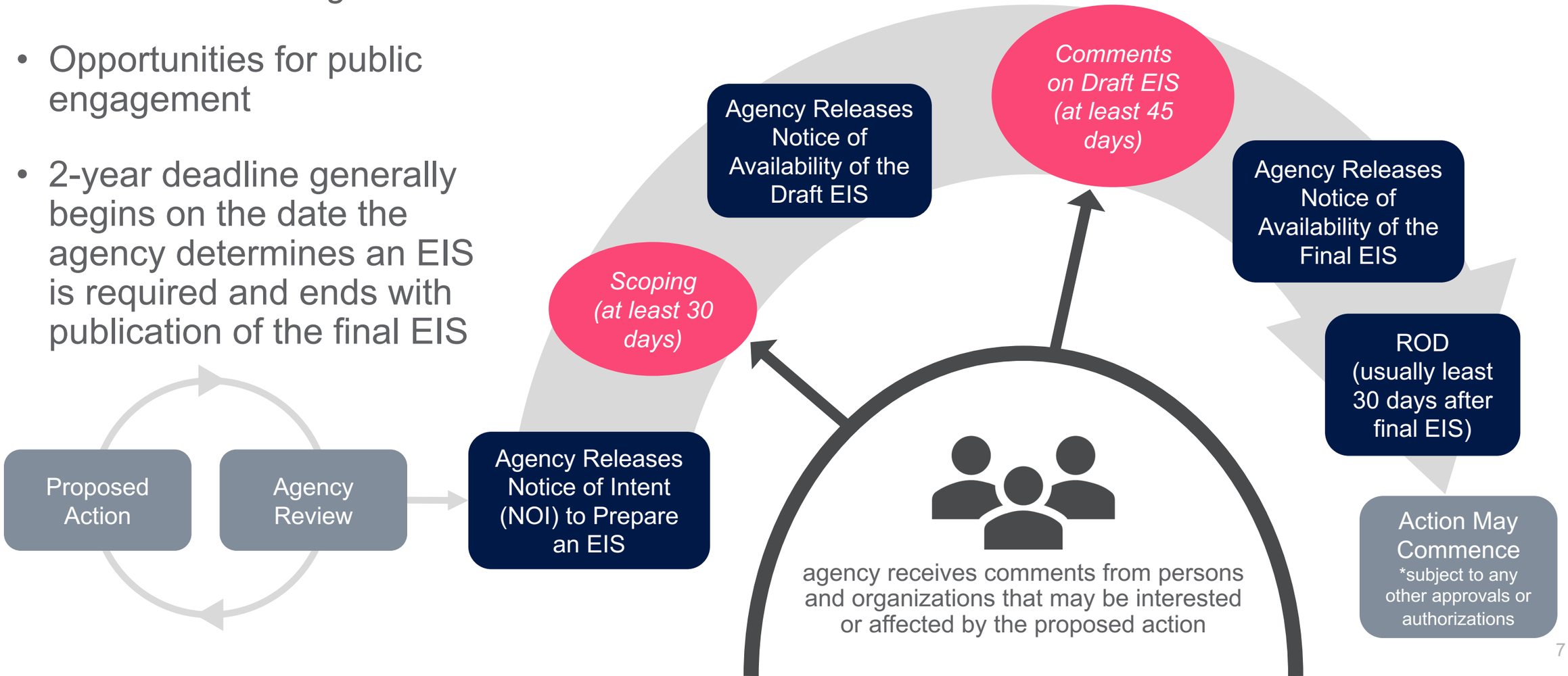
Note: Page limits exclude citations, appendices, and information displayed graphically.

Typical Timelines

- Actual time needed to complete EAs and EISs varies considerably by agency and specific action, but has historically been substantially greater than required under current regulations
- Timelines are influenced by a multitude of factors, including incomplete applications and inadequate funding and staffing of reviewing agencies

EIS Preparation Process

- Major milestones published in the *Federal Register*
- Opportunities for public engagement
- 2-year deadline generally begins on the date the agency determines an EIS is required and ends with publication of the final EIS



Typical Roles

Management and Production

- comment response manager
- document architect (i.e., formatting specialist)
- GIS specialist
- graphic designer
- legal counsel
- NEPA advisor
- project manager
- public engagement specialist
- reference manager
- technical editor

Subject Matter Experts

- air quality and greenhouse gases
- alternatives to the proposed action
- aquatic ecology
- climate change
- cumulative impacts
- environmental justice
- geology, seismology, and soils
- historic and cultural resources
- human health and safety
- hydrology
- land use and visual resources
- meteorology
- noise
- paleontological resources
- socioeconomics
- terrestrial ecology
- transportation and access
- tribal engagement
- waste management

EIS Format

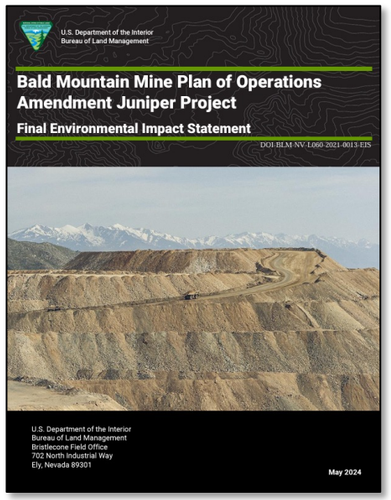
- No universally standardized format, but common elements
- May contain one or more volumes, divided by chapters and appendices
- Predominantly text, but typically contain tables and figures

Example EIS

Final EIS

Document Name	File Size	Release Date	Document Date	Document Type
Dear Interested Party_Juniper-FEIS_signed.pdf	266.77 KB	5/24/2024	5/24/2024	PDF
__Juniper_FEIS_Vol_1.pdf	45.17 MB	5/24/2024	5/24/2024	PDF
App_A_References.pdf	327.85 KB	5/24/2024	5/24/2024	PDF
App_B_Detailed Description.pdf	52.27 MB	5/24/2024	5/24/2024	PDF
App_C_ACEPMs.pdf	378.74 KB	5/24/2024	5/24/2024	PDF
App_D_GRSG Consistency.pdf	346.91 KB	5/24/2024	5/24/2024	PDF
App_E_BMM Eagle Conservation Plan.pdf	15.1 MB	5/24/2024	5/24/2024	PDF
App_F_RFFAs.pdf	388.94 KB	5/24/2024	5/24/2024	PDF
App_G_Supplemental.pdf	2.76 MB	5/24/2024	5/24/2024	PDF
App_H_Glossary.pdf	163.59 KB	5/24/2024	5/24/2024	PDF
App_I_Index.pdf	76.43 KB	5/24/2024	5/24/2024	PDF
App_J_DEIS Comments.pdf	715.13 KB	5/24/2024	5/24/2024	PDF
App_K_FRA_Conformance.pdf	33.98 KB	5/24/2024	5/24/2024	PDF

Often divided into multiple files



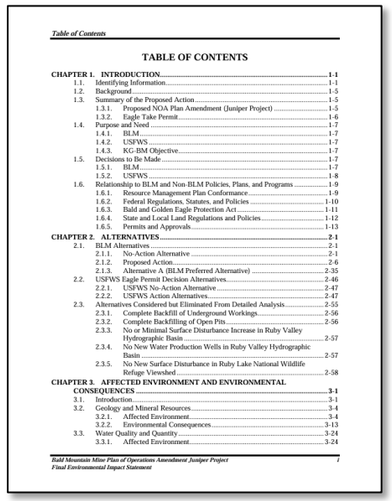
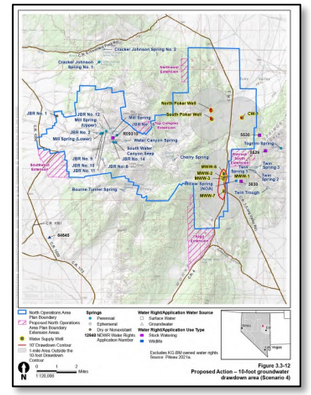


Table of Contents

TABLE OF CONTENTS

Bald Mountain Mine Plan of Operations Amendment Juniper Project *i*
Final Environmental Impact Statement

Organized into chapters and appendices



Information contained in text, figures, tables, and cited references

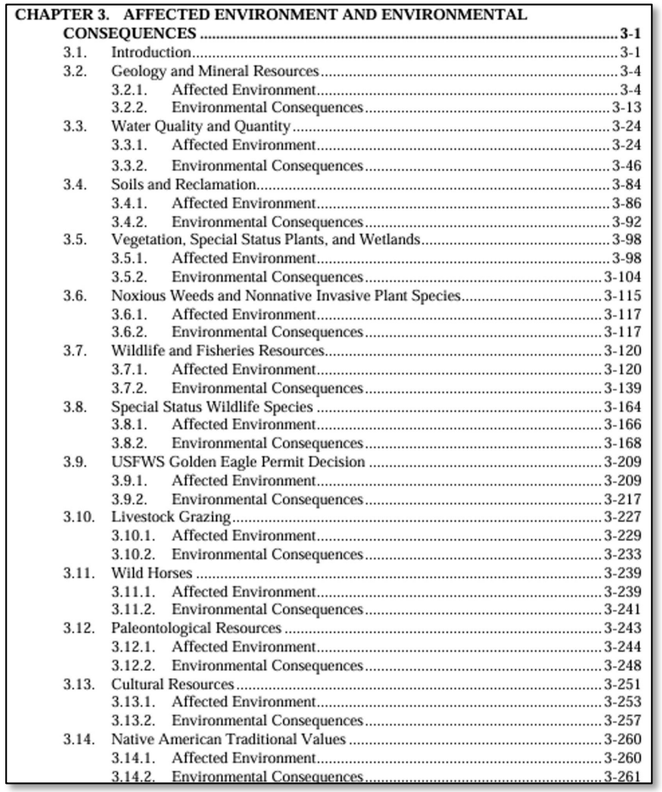
Table 3-8. Comparison of proposed pit floor and groundwater elevations

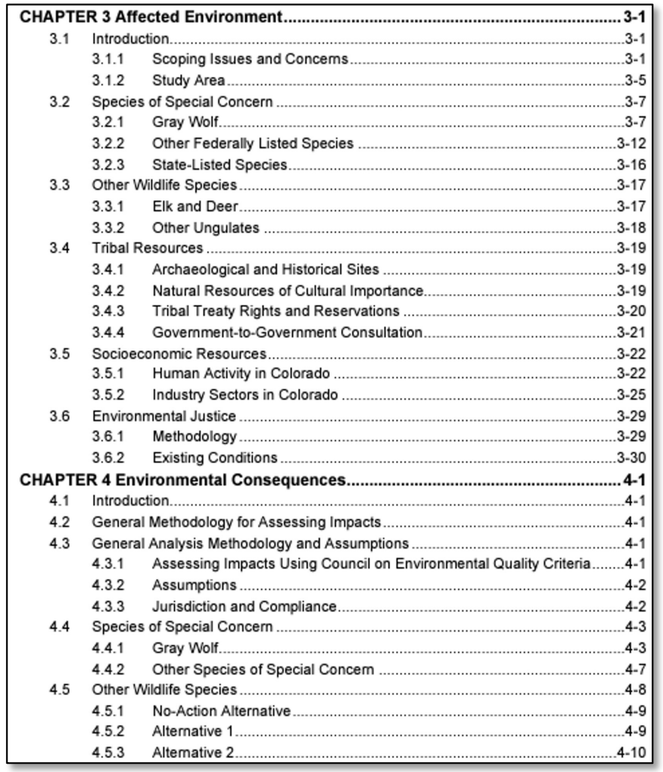
Pit	Floor Elevation (feet amsl) Authorized	Floor Elevation (feet amsl) Proposed	Groundwater Elevation (feet amsl)[1] Minimum	Groundwater Elevation (feet amsl)[1] Maximum	Groundwater Elevation (feet amsl)[1] Average[4]	Estimated Minimum Depth below Pit Floor to Groundwater (feet)[2, 3] Authorized	Estimated Minimum Depth below Pit Floor to Groundwater (feet)[2, 3] Proposed
Redbird	6,620	6,100	5,995	6,054	6,050	566	46
Rat	7,225	6,975	<6,800	<6,800	-	>425	>175
Top expansion	-	7,540	6,702	7,425	6,903	-	115
Royale	-	6,125	6,060	6,095	6,059[5]	-	70[6]
South Duke	-	6,125	6,049	6,069	6,059	-	56
Bida	7,030	6,525	<6,515	6,061	6,061	969	464
Saga	6,225	6,200	5,922	6,143	6,098	82	57[7]
Winrock South	6,600	6,400	6,193	6,556[8]	6,369	216	16[9]

As shown in Table 3-62, the seven tribal communities in the study area had a vacancy rate of 19% in 2020. The South Fork Reservation and Off-Reservation Trust Land had the highest vacancy rate of 44%, and Elko Colony the lowest vacancy rate at 5%. The median home value was 55% lower in the tribal communities than in the study area at $79,400 in 2018. The median rent was 48% lower than in the study area, as well (U.S. Census Bureau 2010b, 2022b, 2022c).

Link to Example EIS: BLM 2024

Variable Format of EISs





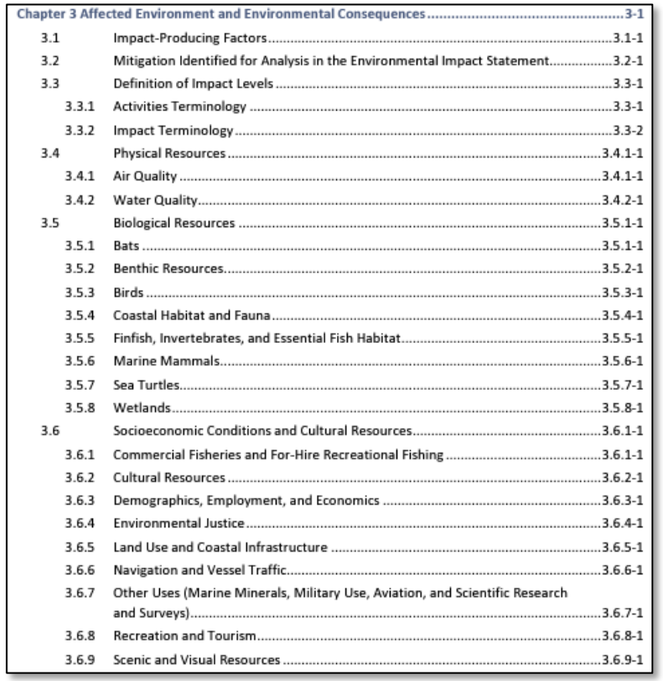
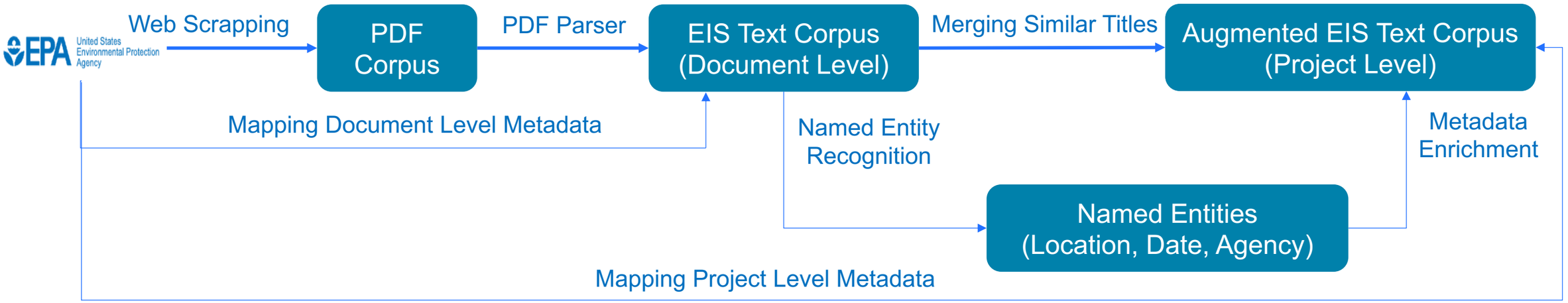


Table of contents of three EISs from different agencies illustrating different scope of resource analyzed and different formats for organizing the discussions of the affected environment and environmental consequences.

EISs pictured (left to right): BLM 2024, USFWS 2023, BOEM 2024

NEPATEC1.0 Construction

NEPA Text Corpus



	Title	Document	EPA Comment Letter Date	Federal Register Date	Agency	State
10	Virginia Reliability Project and Commonwealth ...	Final	-1	09/22/2023	Federal Energy Regulatory Commission	VA
80	Virginia Reliability Project and Commonwealth ...	Draft	06/05/2023	04/21/2023	Federal Energy Regulatory Commission	VA

Exact Duplicates

	Title	Document	EPA Comment Letter Date	Federal Register Date	Agency	State
6802	~ VOIDED ~ Council Bluffs Interstate System Im...	Draft	-1	09/30/2005	Federal Highway Administration	IA
6829	Council Bluffs Interstate System Improvements...	Final	10/07/2005	09/09/2005	Federal Highway Administration	IA
7211	Council Bluffs Interstate System Improvements ...	Draft	01/14/2005	12/23/2004	Federal Highway Administration	IA

Partial Duplicates

NEPATEC1.0 Statistics

- We scrapped a total of **35,427** PDFs from the EPA website from **12,376** EIS Project Links
- We had a total of **16,310** EIS Project Metadata
- We did a 2-step EIS Project Title merging:
  - Duplicate Title Merging
  - Fuzzy Title Merging
- After mapping the PDFs to corresponding metadata, we had a total of **28,212** PDFs from **2,917** Unique EIS Projects
  - Total number of pages: **4.5 Million**
  - Total number of tokens (GPT2 tokenizer): **3.6 Billion**

3K Projects

28K PDFs

4.5M Pages

3.6B Tokens

NEPATEC1.0 Structure

- Dataset is organized by the project, where a project can have multiple documents
- Each project has the following data:
  - Title
  - Project metadata
    - ✓ Agency
    - ✓ States
    - ✓ Dates
  - Page-wise text
    - ✓ NER for each page
    - ✓ Each NER consist of the following features:
      - Text: text for the named entity
      - Label: Label for the named entity
      - Score: Confidence score for the text to belong to the given label

NEPATEC1.0 Named Entity Recognition

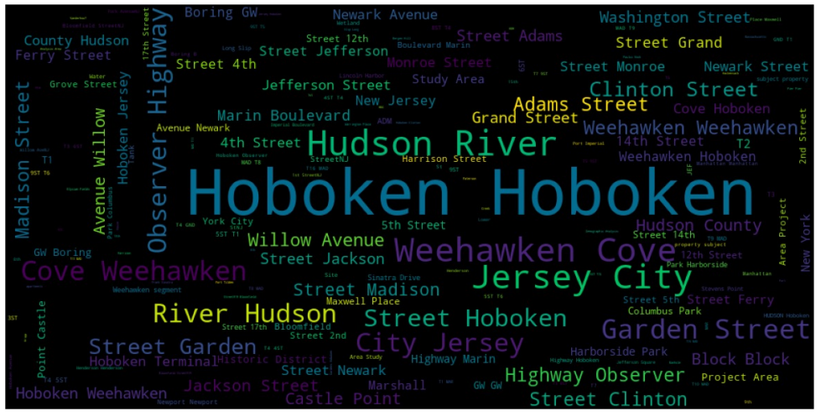
- We extracted a list of five entities from the text:
  - Name: Any name, ranging from name of person to project name
  - Date: Any reference to a specific data or just the year
  - Agency: Any organization
  - Location: Any location, ranging from site location to street, county, state, or country
  - Title: Aimed to extract title of the document and any relevant titles of mentioned documents
- Low threshold to retain major entities
  - Subsample based on score as needed

NORTH CITY Location PROJECT Name EIR EIS RESPONSE TO COMMENTS CAPCOA 2017 Date is the default wind speed for San Diego County Location which is taken from data from the Gillespie Field Location meteorological station Title and includes data from 1996 Date through 2006 Date WRCC 2017 Date This dataset includes hourly wind data as recorded by that station for that time period which includes high wind events Therefore the fugitive dust emissions calculated within CalEEMod Agency account for high wind events within its results From historical records Santa Ana Location winds can easily exceed 50 miles per hour and during a high wind event earth disturbing work would not occur This would be a standard approach by the contractor Name to comply with SDAPCD Agency Rules 55 Fugitive Dust 50 Visible Emissions and 51 Nuisance As stated within the Draft EIR EIS Title the Project Name will comply with all SDAPCD Agency applicable rules Specifically the Project Name would be prevented from allowing emissions during a high wind event by SDAPCD Agency

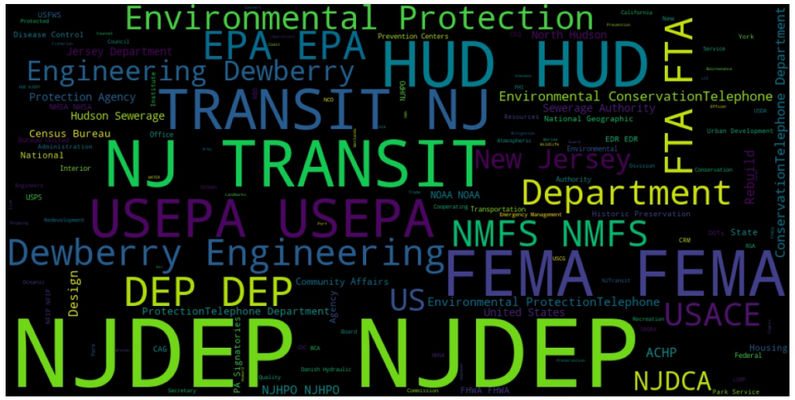
NEPATEC1.0 Named Entity Recognition

Example Word Clouds

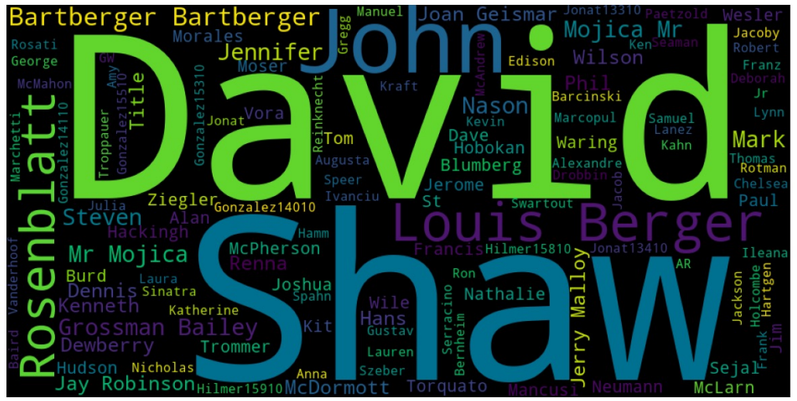
ADOPTION - Rebuild by Design - Hudson River (RBD-HR)



Location Word Cloud



Agency Word Cloud



Name Word Cloud

NEPATEC1.0 Named Entity Recognition

Location NER Tags

ADOPTION - Rebuild by Design - Hudson River (RBD-HR)

```
['Bloomfield StreetNJ',
 '12th Street',
 'Park Avenue',
 'Hoboken Avenue',
 'Paterson Avenue',
 'Adams StNJ',
 'Jefferson Street',
 'GW 5',
 'Willow Avenue',
 'Vezzetti Way',
 'Jersey Avenue',
 'Willow AveNJ',
 '6th Street',
 'Shippen Street',
 'Park Avenue',
 'Clinton Street',
 '8th Street',
 'Frank Sinatra Drive',
 '16th Street',
 'Newark Street',
 'Jefferson Street',
 'Newark Avenue',
```

Street Locations

```
['Hudson county', 'Hudson County']
```

County Locations

```
['NJ', 'Massachusetts', 'NJ11', 'New Jersey', 'NJ']
```

State Locations

```
['Weehawken Cove',
 'Newark',
 'Stevens Point',
 'Princeton',
 'Castle Point',
 'Warrington Plaza',
 'Hoboken',
 'Winfield',
 'Hudson Falls',
 'Englewood Cliffs',
 'Hudson',
 'Newport',
 'Hoboken',
 'Hoboken',
 'New York City',
 'Fort Lee',
 'Manhattan',
 'Madison',
 'Henderson',
 'Philadelphia',
 'Weehawken',
 'Vanderhoof',
 'Hudson PlaceNJ',
 'NYC',
 'San Francisco']
```

City Locations

NEPATEC1.0 Named Entity Recognition

Location NER Tags

ADOPTION - Rebuild by Design - Hudson River (RBD-HR)

Reference Section/Document

Challenges and Opportunities

- Existing online repositories containing EISs and other NEPA documents are incomplete and have limited search capabilities
- EISs are long documents that do not have a universally consistent or standardized structure
- AI can assist people involved in the preparation and review of NEPA documents and enhance efficiency by:
  - searching for, interpreting, and synthesizing content from existing NEPA documents
  - assessing trends in NEPA documents and processes over time
  - generating draft content or reviewing text of new NEPA documents
  - assisting with analysis and summarization of public comments

Technical Opportunities, Users, and Solutions

Below we present a set of problems, ways solutions would benefit potential users (high-level stakeholders, researchers/authors of EISs, and members of the public), and examples of potential solutions.

- **NEPA documents often contain technical jargon and are challenging for the public to understand.**
  - Need: Apply language modeling to suggest revisions to NEPA documents to improve clarify through use of plain text (e.g., clear, understandable prose)
  - Solutions:
    - ✓ Develop AI-powered tools for enhancing document clarity and readability.
    - ✓ Implement a plain language checker and recommendation for NEPA documents.

- **High-Level Stakeholders:** Ensures documents meet policy and regulatory standards.
- **Researchers:** Helps in producing clear and concise research outputs.
- **General Public:** Improves document readability and comprehension.

Technical Complexity Level

- **Reviewing and synthesizing information from multiple NEPA documents is time consuming.**
  - Need: Enable rapid comparative analysis and summarization of multiple NEPA documents.
  - Solutions:
    - ✓ Develop natural language processing algorithms to summarize documents.
    - ✓ Create a comparative analysis tool for multiple NEPA documents.

- **High-Level Stakeholders:** Provides quick insights for decision-making.
- **Researchers:** Assists in comparative studies across multiple documents.
- **General Public:** Simplifies complex information for better understanding.

Technical Complexity Level

Technical Opportunities, Users, and Solutions (Continued)

- **NEPA documents are information rich but have limited to no metadata.**
  - Need: Extract multi-level metadata to facilitate information retrieval and analysis by LLMs.
  - Solutions:
    - ✓ Develop software to reliably harvest metadata from a variety of file types (PDFs, Word documents, etc.)
    - ✓ Implement a metadata standardization system to ensure consistency across data sets.

- **High-Level Stakeholders:** Helps in policy-making and oversight.
- **Researchers:** Aids in academic and practical research.
- **General Public:** Assists in understanding project context and relevance.

Technical Complexity Level

- **How do NEPA projects and processes change with place and time?**
  - Need: Find documents or assess spatio-temporal trends based on project location and points in time.
  - Solutions:
    - ✓ Create visualization tools for spatio-temporal data analysis.
    - ✓ Develop machine learning models to predict trends based on historical data.

- **High-Level Stakeholders:** Enables strategic planning based on spatial and temporal trends.
- **Researchers:** Aids in identifying patterns and anomalies over time.
- **General Public:** Enhances public awareness of environmental changes.

Technical Complexity Level

Technical Opportunities, Users, and Solutions (Continued)

- **NEPA documents are mostly text and not designed to convey information quickly or visually.**
  - Need: Display specific subsets of information graphically to enhance understanding (e.g., word cloud, complex numerical taxonomic clustering).
  - Solutions:
    - ✓ Develop dynamic visual tools like word clouds and taxonomic clustering.
    - ✓ Create dashboards for visual representation of complex data sets.

- **High-Level Stakeholders:** Assists in data-driven decision-making.
- **Researchers:** Helps to visually interpret complex data patterns.
- **General Public:** Makes technical information accessible through visuals.

Technical Complexity Level

- **Few quantitative metrics are available to measure the efficiency of the NEPA process.**
  - Need: Assess trends in document length, process length, project types, and other characteristics over time.
  - Solutions:
    - ✓ Develop tools to track document length, process duration, and other efficiency metrics.
    - ✓ Implement dashboards for real-time monitoring of NEPA efficiency.

- **High-Level Stakeholders:** Informs policy reviews and legislative changes.
- **Researchers:** Supports empirical studies and trend analysis.
- **General Public:** Offers transparency in governmental processes.

Technical Complexity Level

Technical Opportunities, Users, and Solutions (Continued)

- **Geospatial information is not directly embedded in NEPA documents.**
  - Need: Extract detailed project location data to enable geographical search and localization capabilities.
  - Solutions:
    - ✓ Create a geographic information system (GIS) integration module.
    - ✓ Develop APIs to extract and standardize location data from NEPA documents.

  - **High-Level Stakeholders:** Enhances geographical planning and resource allocation.
  - **Researchers:** Facilitates spatial analysis and correlation studies.
  - **General Public:** Provides an easy way to identify projects in their vicinity.

Technical Complexity Level

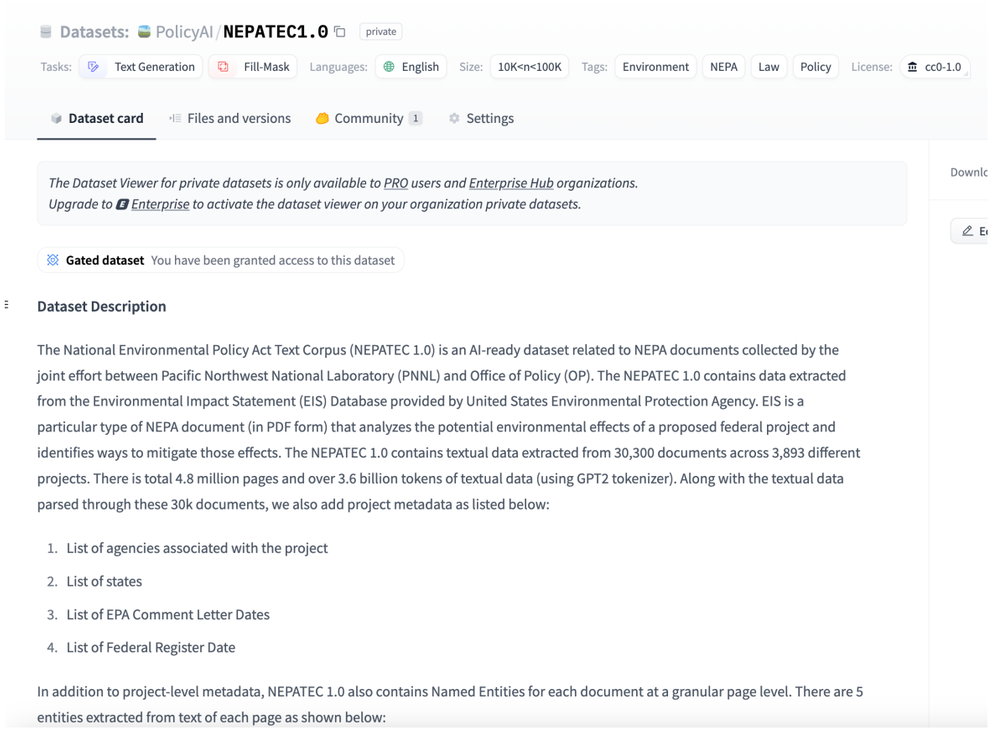
- **Are NEPA documents using best available science?**
  - Need: Identify and assess trends in scientific concepts and studies cited in NEPA documents over time and space
  - Solutions:
    - ✓ Develop text mining and analysis tools to identify key scientific concepts.
    - ✓ Implement trend analysis modules for scientific studies cited in NEPA documents.

  - **High-Level Stakeholders:** Informs evidence-based policy-making.
  - **Researchers:** Enables trend analysis in scientific research.
  - **General Public:** Promotes awareness of the scientific foundations of projects.

Technical Complexity Level

NEPATEC1.0 Dataset

- NEPATEC1.0 Dataset is publicly available in HuggingFace under Creative Commons “0” license (public domain dedication)



Datasets: PolicyAI / **NEPATEC1.0** private

Tasks: Text Generation | Fill-Mask | Languages: English | Size: 10K<n<100K | Tags: Environment | NEPA | Law | Policy | License: cc0-1.0

Dataset card | Files and versions | Community 1 | Settings

The Dataset Viewer for private datasets is only available to PRO users and Enterprise Hub organizations.
Upgrade to Enterprise to activate the dataset viewer on your organization private datasets.

Gated dataset You have been granted access to this dataset

Dataset Description

The National Environmental Policy Act Text Corpus (NEPATEC 1.0) is an AI-ready dataset related to NEPA documents collected by the joint effort between Pacific Northwest National Laboratory (PNNL) and Office of Policy (OP). The NEPATEC 1.0 contains data extracted from the Environmental Impact Statement (EIS) Database provided by United States Environmental Protection Agency. EIS is a particular type of NEPA document (in PDF form) that analyzes the potential environmental effects of a proposed federal project and identifies ways to mitigate those effects. The NEPATEC 1.0 contains textual data extracted from 30,300 documents across 3,893 different projects. There is total 4.8 million pages and over 3.6 billion tokens of textual data (using GPT2 tokenizer). Along with the textual data parsed through these 30k documents, we also add project metadata as listed below:

1. List of agencies associated with the project
2. List of states
3. List of EPA Comment Letter Dates
4. List of Federal Register Date

In addition to project-level metadata, NEPATEC 1.0 also contains Named Entities for each document at a granular page level. There are 5 entities extracted from text of each page as shown below:

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

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 32,000 EIS documents associated with 3,893 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, and agency must first prepare a more concise document called an environmental assessment (EA) to support its determination (Figure 1).

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. 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) [1]. An agency typically begins the NEPA process after determining

Preprint. Under review.

Community Outreach

Generative AI Challenge for Environmental Review

Kaggle competition: “**LLM for Environmental Review**"

Dataset: NEPAQuAD1.0

Task: Develop LLM for Question Answering

Start Date: **May 15, 2024.**

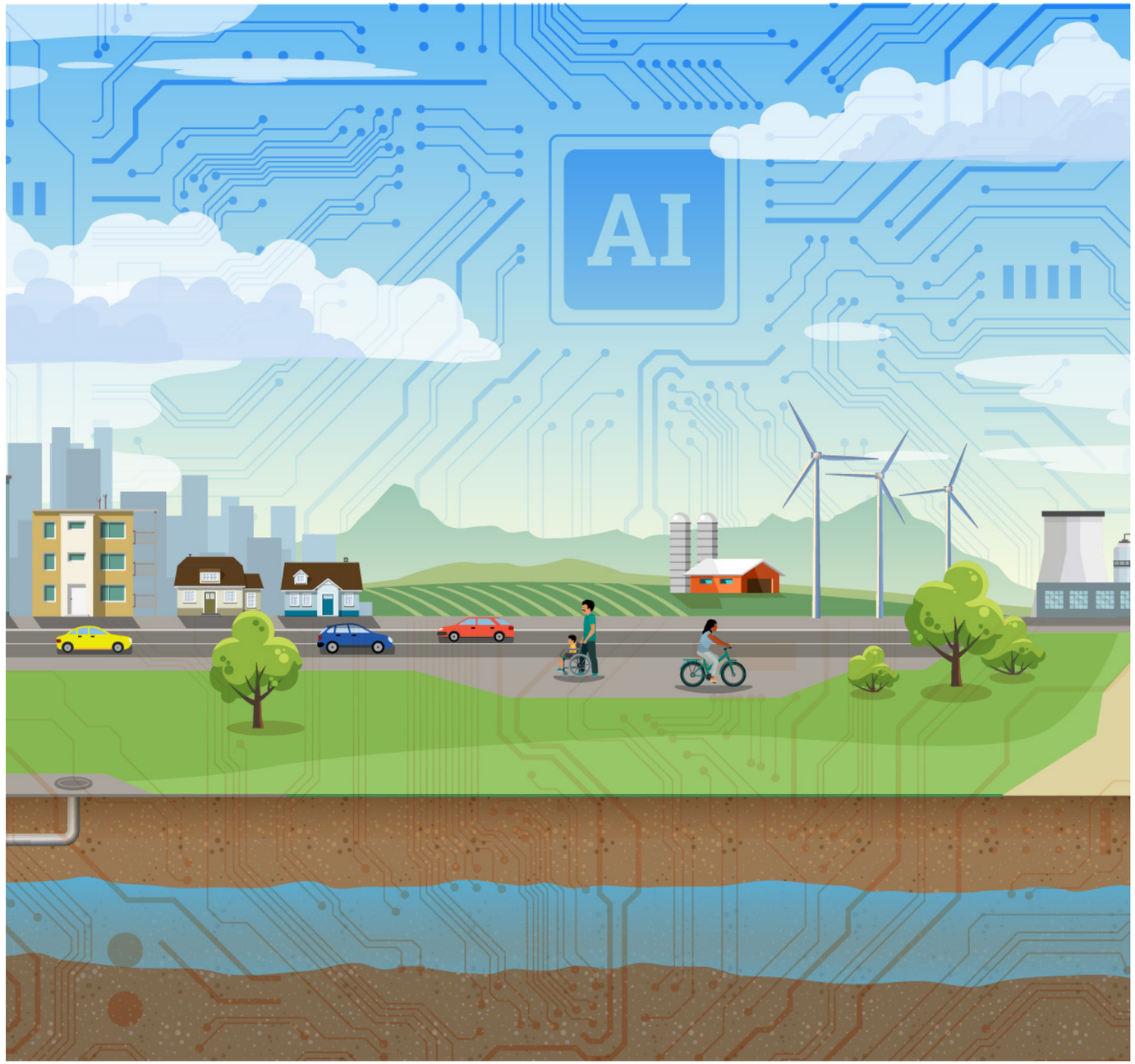
End Date: June 30, 2024

PolicyAI

Pacific Northwest
NATIONAL LABORATORY

Acknowledgement

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.



Pacific Northwest
NATIONAL LABORATORY

Thank you