# NEPATEC v2.0: Standardized Metadata and Text Corpus of National Environmental Policy Act Documents

**Sai Munikoti[1,+,*], Dan Nally[1,+], Sai Dileep Koneru[1,+], Siddhartha Shankar Das[1,+], Kaustav Bhattacharjee[1,+], Ashik Islam[1,+], Alex Buchko[1], Taylor Edwards[1], Kathy Nwe[1], Siddhisanket Raskar[1], Paul Rigor[1], Heng Wan[1], Micah Taylor[1], Scott Spare[1], Derek Lilienthal[1], Mahantesh Halappanavar[1], Anurag Acharya[1], Tim Vega[1], Mike Parker[1], Anastasia Bernat[1], and Sameera Horawalavithana[1,+,*]**

[1]Pacific Northwest National Laboratory, Richland, WA
[*]permitai@pnnl.gov
[+]these authors contributed equally to this work

## ABSTRACT

The National Environmental Policy Act of 1969, as amended (NEPA), is a major environmental law in the United States, requiring Federal agencies to consider and document potential environmental impacts before deciding on a proposed action. Modernization of NEPA and permitting processes faces significant challenges due to the lack of standardized formats and interoperable systems for organizing and sharing NEPA-related information across agencies. Much of the information gathered during NEPA reviews is written into documents such as categorical exclusions, environmental assessments, and environmental impact statements, then filed in predominately independent agency file stores that may or may not be publicly accessible. The application of metadata and data standards, such as those recommended by the Council on Environmental Quality (CEQ), to NEPA documents offers a shared vocabulary and structure for key entities like projects, processes, and documents that can streamline information exchange and enhance collaboration across systems. In this work, we publicly release NEPATEC2.0, an expanded corpus of public NEPA documents with associated metadata. NEPATEC2.0 consists of more than 120,000 documents from 60,000 projects prepared by more than 60 different agencies. Modeled to align with CEQ metadata standards, NEPATEC2.0 promotes consistency in environmental reviews and supports the ongoing effort to modernize permitting technologies by facilitating more transparent, efficient, and data-driven decision-making. The NEPATECv2 Dataset is publicly accessible at https://huggingface.co/datasets/PNNL/NEPATEC2.0.

## 1 Background & Summary

The National Environmental Policy Act of 1969, as amended (NEPA), is a bedrock environmental law in the United States that has also been replicated internationally. The express intent of NEPA is fostering a productive harmony between humans and the environment for present and future generations. The NEPA statute (42 U.S. Code 4321 et seq.) requires federal agencies to consider reasonably foreseeable environmental effects and potential alternatives in their decisions on agency actions. Federal agencies must first determine whether NEPA applies to a proposed action and then determine the appropriate level of environmental review. A categorical exclusion (CE) is the most basic level of NEPA review and addresses those categories of actions that a Federal agency has determined do not normally have a significant effect on the environment and are therefore categorically excluded from detailed environmental review. 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.

Each type of NEPA review requires preparation of a written document disclosing relevant information that supports the agency's decision-making process. The NEPA statute now limits EAs to 75 pages and, in most cases, limits EISs to 150 pages, excluding citations, appendices, and information displayed graphically. Historically, many EISs have been substantially longer than 150 pages. Average document length for EISs sampled by the CEQ from 2013 to 2018 was 575 pages for draft documents and 661 pages for final documents[1].

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.

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 related application requirements are met. Before preparing an EIS, agencies publish a notice of intent in the Federal Register and accept public comments, usually for a period of 30 days. Historically, agencies have released draft and final versions of each EIS. Less commonly, in event a proposed action or environmental circumstances change considerably, agencies may prepare a supplement to a draft or final EIS. With the removal of CEQ's regulations implementing NEPA in 2025, there is no longer a statutory requirement to release a draft EIS for public comment. At the conclusion of a NEPA process requiring an EIS, an agency signs a record of decision stating its decision, identifying the alternatives analyzed, and any required mitigation. EISs often have various appendices as well as 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. The process for preparing an EA is similar to, but shorter than an EIS, and does not require publication of Federal Register notices or public scoping. The agency decision is documented in a "finding of no significant impact" rather than a record of decision.

Metadata standards are essential for modernizing NEPA and permitting systems and agencies have been directed to work towards the service delivery standards and minimal functional requirements outlined in the CEQ's Permitting and Technology Action Plan.[2] NEPA documents are currently stored in various agency database and metadata, if present, is typically minimal and does not follow a standard format. Large language models (LLMs) and artificial intelligence (AI) offer potential solutions to retroactively extract and compile metadata from a large number of completed NEPA documents, which may otherwise be prohibitively expensive or time consuming to produce manually. However, previous works have identified the challenges in cataloging NEPA documents, particularly EISs, through automated techniques[3,4]. 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. Additionally, agency databases often contain variable other types of related or supporting documents that must be distinguished from NEPA documents.

In this work, we publicly release a large text-corpus of NEPA permitting documents, named as **N**ational **E**nvironmental **P**olicy **A**ct **Te**xt **C**orpus (NEPATEC2.0). NEPATEC2.0[1] improves the previous version of the NEPA public text corpus, NEPATEC1.0,[2] as well as draws inspiration from previous works. In comparison to NEPATEC1.0, NEPATEC2.0 contains more than 120K number of documents and 14 number of metadata that align with the metadata standards recommended by the CEQ, who defines concepts, categories, and relationships within the environmental review domain (see Table 1 for additional data statistics). These standards promote a shared vocabulary for organizing and standardizing entities like projects, processes, and documents. Agencies can use these standards to share a common digital language, improving interoperability and streamlining data exchange across agency systems.[5]

Navigating through thousands of NEPA documents as individual files or extracted text is challenging both for human users as well as AI models. Layering metadata on top of the documents can aid both human users and AI models in contextualizing the search space and facilitate the seamless, accurate, and reliable retrieval of required information. NEPATEC2.0 advances the organization of the historical data into a recommended standard form, which will facilitate more targeted and rapid search and query of historical documents. Historical documents can provide examples and context for alternatives development, environmental analysis, regulatory approaches, and public concerns. They can also be used to identify high-level trends in agency NEPA reviews based on attributes such as project type and location.

## 2 Methods

### 2.1 NEPA Document Collection and Preprocessing

Table 1 describes the NEPATEC document collection. In contrast to NEPATEC1.0 that mainly consists of EIS documents from the Environmental Protection Agency (EPA's) NEPA Compliance Database,[3] we expanded our data sources for NEPATEC2.0 to include three additional federal agencies such as Department of Energy (DOE), Department of Agriculture (USDA), and

[1] https://huggingface.co/datasets/PNNL/NEPATEC2.0
[2] https://huggingface.co/datasets/PNNL/NEPATEC1.0
[3] https://www.epa.gov/nepa/epa-compliance-national-environmental-policy-act

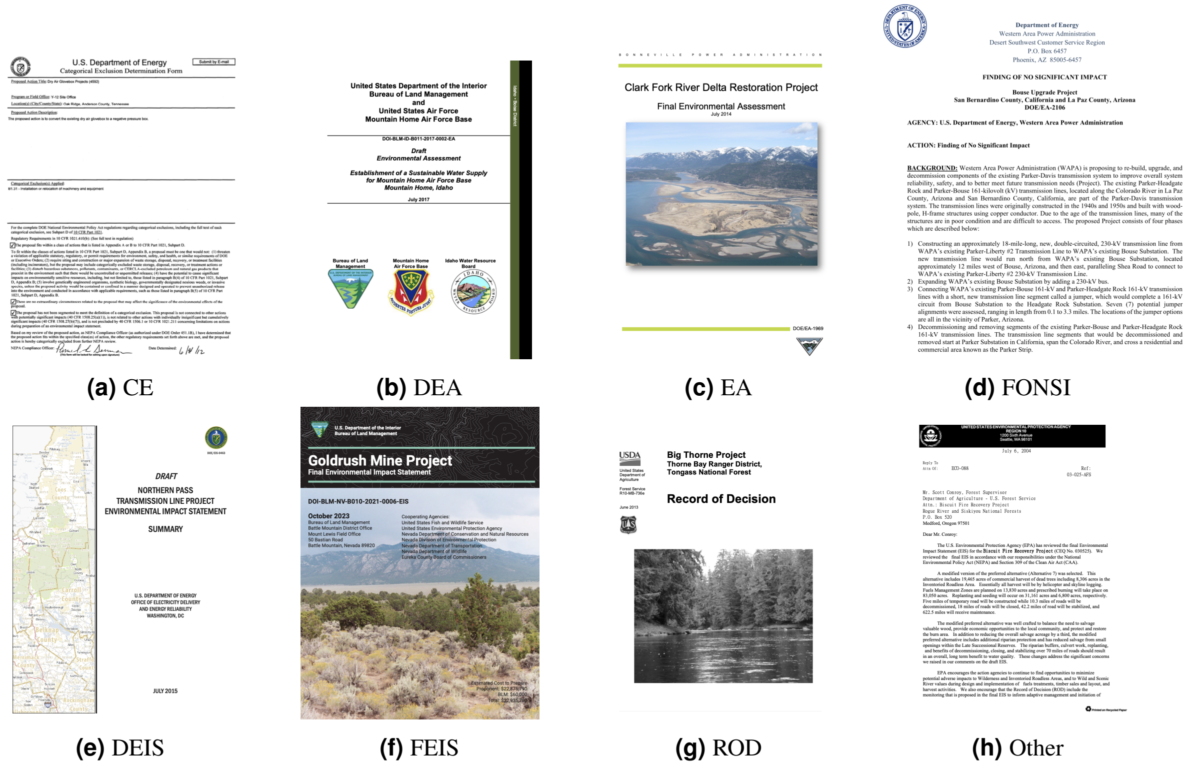
| Data Provider | Intake Medium | Document Type | #Files | #Pages | NEPATEC Version (Public) | Cut-Off Date |
|---|---|---|---|---|---|---|
| EPA | Web Crawling, and Scraping | EIS | 27,648 | 3,989,947 | 1.0 | 11/2023 |
| | API | EIS | 3020 | 584,877 | 2.0 | 07/2025 |
| DOE | Web Scraping | CE | 31,377 | 65,850 | 2.0 | 10/2023 |
| | | EA | 2354 | 170,251 | 2.0 | 10/2023 |
| | | EIS | 2410 | 431,579 | 2.0 | 10/2023 |
| USDA | Zip via PNNL File Transfer | CE | 177 | 1,396 | 2.0 | 02/2024 |
| | | EA | 33 | 2064 | 2.0 | 02/2024 |
| BLM | Web Scraping | CE | 43,793 | 311,075 | 2.0 | 05/2024 |
| | | EA | 11,855 | 296,791 | 2.0 | 05/2024 |
| | | EIS | 21,219 | 1,125,354 | 2.0 | 05/2024 |
| | | **Total** | **143,886** | **6,979,184** | | |

**Table 1.** NEPATEC Data Collection. For EIS obtained from the EPA website, we iteratively find and fetch web links starting from a list of seed NEPA document download URLs. We download DOE and BLM documents from agency-provided document URLs.

the Bureau of Land Management (BLM) under the Department of Interior. Note that, the majority of USDA documents were from Forest Service (FS). These agencies account for a large percentage of the total number of federal NEPA reviews. There is currently no one-size-fits-all solution for aggregating NEPA documents from federal agencies, as each agency stores its documents in different systems (such as bulk PDF storage or structured databases) and offers different access points and metadata. To effectively collect these documents, we leverage a variety of methods, including web scraping from websites/databases, programmatic pulling from application programming interfaces (APIs), and direct file transfers via email, among others. We are grateful for the support and cooperation of the agencies in facilitating our collection of these public-domain documents.

NEPATEC2.0 contains the following document types:

- Categorical Exclusion (CE): Administrative determination that exempts a proposed Federal action from detailed environmental review if it falls within a category of activities that the agency has determined do not have significant environmental effects. These exclusions streamline the NEPA process for routine actions with well-established and minimal environmental impacts.
- Draft Environmental Assessment (DEA): Preliminary analysis document that evaluates the potential environmental consequences of a proposed federal action to determine whether significant impacts demands preparation of a full environmental impact statement. This assessment serves as the initial comprehensive environmental review for actions/projects of uncertain environmental significance. Agencies have the option to release a draft environmental assessment but are not required to.
- Final Environmental Assessment (FEA): Final version of an environmental assessment before the agency signs a finding of no significant impact. If an agency releases only a single EA prior to its finding, we label it as "FEA."
- Finding of no significant impact (FONSI): Official determination that concludes a proposed federal action will not result in significant environmental effects, thereby eliminating the need to prepare an EIS. A FONSI is the final decision document concluding an EA-level NEPA review.
- Draft Environmental Impact Statement (DEIS): First version of an EIS document shared with external parties. An EIS provides detailed analysis of the environmental consequences of proposed major federal actions that could have significant effects on the environment and considers potential alternatives to the proposed action and public comments received during the scoping period.
- Final Environmental Impact Statement (FEIS): Final publicly released version of an EIS document before the agency signs a record of decision.
- Record of Decision (ROD): Official decision document that identifies the selected alternative from those analyzed in the EIS, including the rationale for alternative selection and any mitigation measures or monitoring requirements. A ROD is the final decision document concluding an EIS-level NEPA review.


U.S. Department of Energy
Categorical Exclusion Determination Form
United States Department of the Interior
Bureau of Land Management
and
United States Air Force
Mountain Home Air Force Base
Draft
Environmental Assessment
Establishment of a Sustainable Water Supply for Mountain Home Air Force Base
Mountain Home, Idaho
July 2017
Clark Fork River Delta Restoration Project
Final Environmental Assessment
July 2014
Department of Energy
Western Area Power Administration
FINDING OF NO SIGNIFICANT IMPACT
Bouse Upgrade Project
DRAFT
NORTHERN PASS
TRANSMISSION LINE PROJECT
ENVIRONMENTAL IMPACT STATEMENT
SUMMARY
JULY 2015
Goldrush Mine Project
Final Environmental Impact Statement
DOI-BLM-NV-B010-2021-0006-EIS
October 2023
Big Thorne Project
Thorne Bay Ranger District,
Tongass National Forest
Record of Decision
UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

**(a)** CE

**(b)** DEA

**(c)** EA

**(d)** FONSI

**(e)** DEIS

**(f)** FEIS

**(g)** ROD

**(h)** Other

**Figure 1.** Cover pages of various document types

- Other supporting documents (Other): Supplementary analytical and administrative documents that provide additional technical detail, related procedural information, or specialized analyses beyond the primary NEPA documents. This category encompasses NEPA document appendices containing detailed technical data, addendums providing clarifications or corrections, public comments, and or any other document that does not fit one of the seven categories above.

Figure 1 depicts the cover page of various documents types in NEPATEC2.0. The length of documents varies by NEPA review type, with CE generally being the shortest and EIS being the longest and most rigorous.

## 2.2 NEPA Metadata Identification

An important step in modernizing NEPA and permitting systems is progressing toward a government-wide data and technology standard for NEPA that provides agencies with a common digital language to facilitate interoperability and automatic data exchange among systems.[6] CEQ has recommended an initial data and technology standard as part of the Permitting Technology Action Plan.[2] This standard outlines methods to organize and standardize various concepts/entities (e.g., projects, processes, and documents) that are generally involved in defining the NEPA or permitting review processes and how these entities relate to one another.

- **Project** represents an agency's activity or decision requiring initiation of one or more related processes.
- **Process** refers to the environmental review, permit, or authorization that an agency is conducting, issuing, or otherwise completing as part of the decision-making process.
- **Documents** are created during the environmental review, permit, or authorization process.
- **Files** are an electronic storage medium for documents. A document may consist of on or more files, such as an EIS document divided into volumes.

A "Project" entity may involve one or more related "Process" entities (e.g., review under NEPA, construction permit, or interagency consultation). Each "Process" entity may involve the creation or issuance of "Documents" (e.g., Notice of Intent), which may consist of one or more "Files."

We worked with CEQ to identify a subset of project, process, and document properties (referred to as metadata here onwards)

**Table 2.** Metadata attributes grouped by entity. All attributes are available across EIS, EA, and CE process types, except those marked with an asterisk (*), which are available only for CE processes.

| Entity | Metadata Attribute | Description | Datatype |
|---|---|---|---|
| **Process** | lead_agency | Federal or other agency responsible for conducting the process. | text |
| | process_family | Major category that the process belongs to. | text |
| | process_type | Type of review or permitting process. A sub-type of process family. | text |
| **Project** | project_title | Descriptive name of the project. | text |
| | location | Name of city, county, or other geographic area where the project is located. | text |
| | project_sponsor | Name of responsible entity, organization, or person for project. | text |
| | project_sector | High-level project sector(s). | text |
| | project_type | Type(s) of project. A sub-type of project sector. See list in Table 10. Categorization is based on best fit and is subjective. | text |
| **Document** | document_type | Type of document. | text |
| | document_title | Title of document, reflecting the actual title, not the file name. | text |
| | prepared_by | Agency or entity (e.g., contractor) responsible for preparation. | text |
| | ce_category* | Specific category or categories under which the action is classified as a CE. | text |
| | action_description* | Brief summary of the proposed action, including its purpose and need. | text |
| **File** | section_or_volume_title | Title of a specific document section or volume. | text |
| | main_document | Indicates if file is the main document ("Yes") or supporting info ("No"). Main document consists of the title page and executive summary through all chapters, but excludes appendices. | boolean |

that have fundamental importance for document classification and discovery and that can generally be extracted directly from historical NEPA documents. For our purposes, we added several file-level metadata properties to account for the inherited structure of documents that were divided into multiple files.

For NEPATEC2.0, we have identified 14 essential metadata properties and developed a AI-driven framework for extracting them from documents. The definition of each metadata are tabulated in Table 2.

## 2.3 NEPA Metadata Extraction Framework

We developed a *NEPA Metadata Extraction Framework* with a systematic workflow designed to transform unstructured environmental documents into structured, analyzable data (Figure 2). This framework consists of four distinct stages: it begins with **documents intake**, where raw files are collected and ingested. Next, **data pre-processing** (e.g., PDF to text and image parsing) cleans and standardizes this information to make it machine-readable. During **metadata enrichment** stage, we developed advanced Natural Language Processing (NLP) algorithms to identify and extract key metadata. Finally, the **NEPATEC schema updates** stage involves loading this newly structured extracted information into a central database.

Figure 3 describes the NLP methods of extracting file, document, process and project level metadata from text documents. Our metadata extraction methods rely on LLM-based prompting methodology that leverages state-of-the-art NLP capabilities. The framework uses *Gemini 2.5 Flash*[7] as the core LLM, selected specifically for its extensive context length ($1M$) capacity that enables the accommodation and processing of lengthy NEPA decision documents without truncation. Further, the framework incorporates structured output generation capabilities through Pydantic, ensuring that extracted metadata adheres to predefined data type (string, numeric, categorical, etc.) and maintains consistency across different document types. Most importantly, the extraction framework implements automatic optimization of seed prompts through DSPy,[8] which improves the initial prompt versions developed by subject matter experts (SMEs). In this regard, we first curated a subset of around 300 documents, each

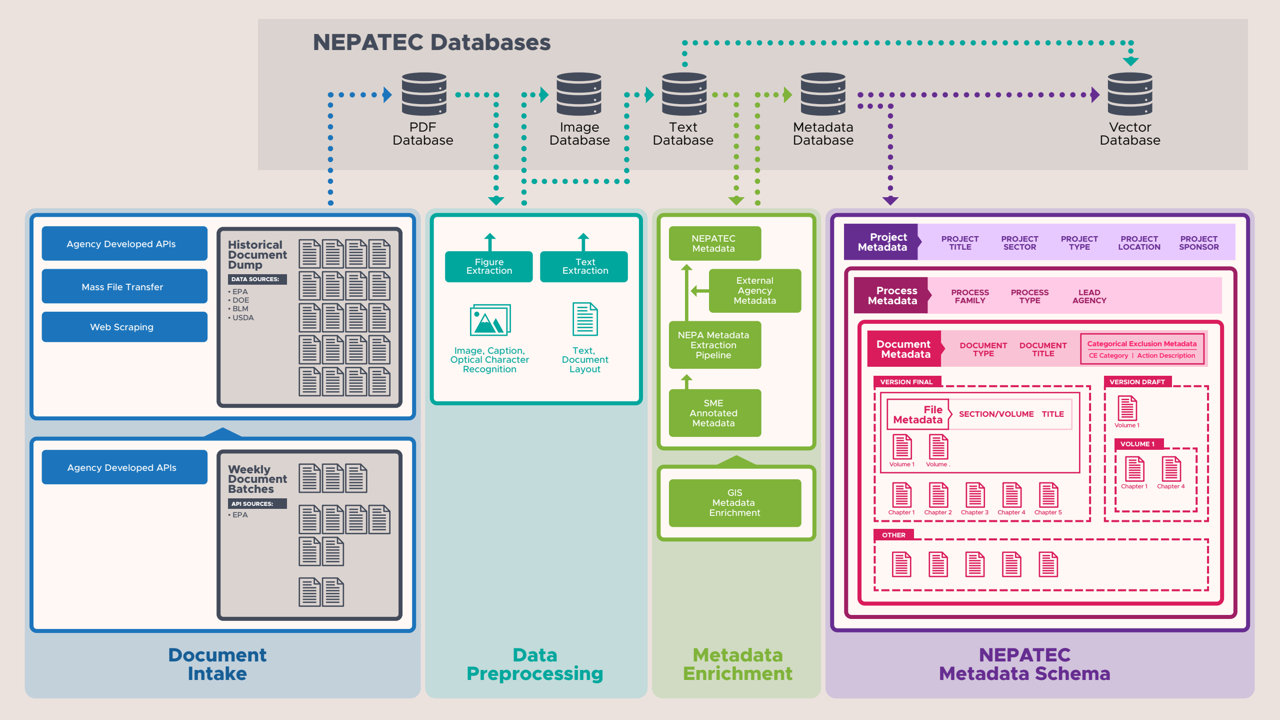


**Figure 2.** NEPA Metadata Extraction Framework consists of documents intake, data pre-processing, metadata enrichment and NEPATEC schema updates stages.

.

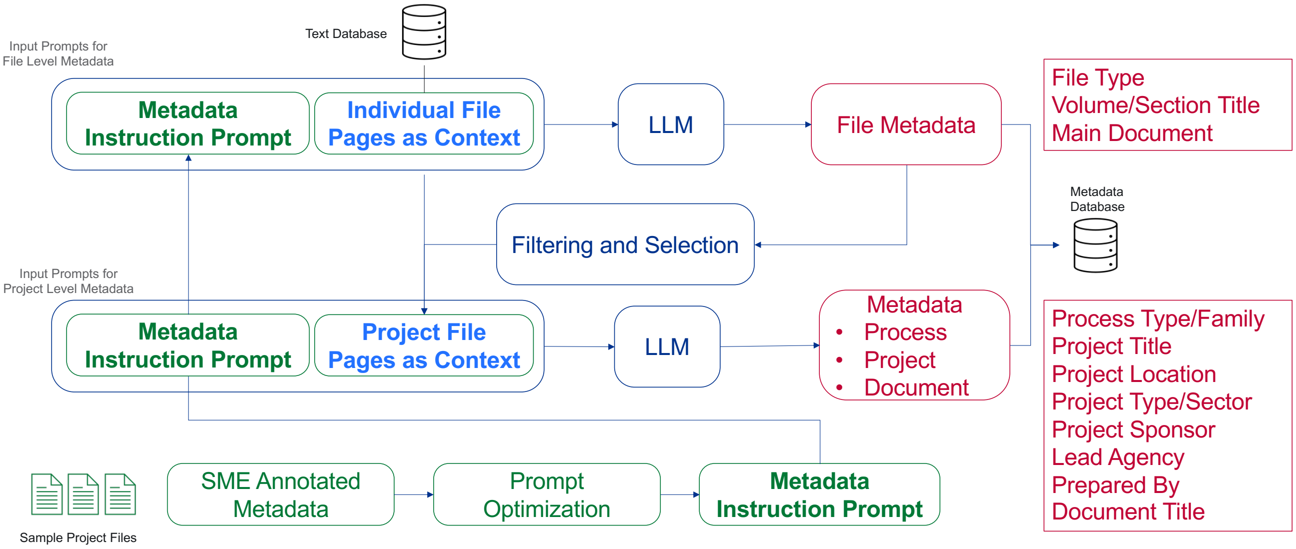


**Figure 3.** AI-Driven NEPA Metadata Extraction Pipeline.

annotated by SMEs with 10 to 14 metadata attributes, depending on the type of document. This curated dataset serves as the ground truth for prompt optimization and method validation. During the optimization process, the input string and the required context (first-k pages from the beginning) from the documents are modulated based on their performance on the annotated set. This process yields a customized prompt and the necessary context for each type of metadata and document, which are used to scale the extraction across all documents in the NEPATEC2.0 dataset. It is important to understand that the extraction of metadata can vary depending on the number of files required. Some metadata types are extracted from a single file (i.e., requires input text from single file), while others require multiple files (requires input text from multiple files). For example, *document_type, section/volume_title, and main _document* are based on single file level extraction. On the other hand, extraction of project metadata may be based on multiple files (see Figure 3).

Overall, this metadata extraction framework demonstrates the potential utility of automated extraction methods to organize and standardize metadata from NEPA documents in accordance with CEQ's recommendations and to improve data accessibility.

### 2.4 Cloud Implementation

We leverage two primary services within the Google Cloud Platform (GCP) to deploy the metadata enrichment methods: *Google Cloud Workflows* and *Batch Prediction with Gemini*.

*Google Cloud Workflows* is a fully managed orchestration service within the Google Cloud Platform (GCP) ecosystem that enables the definition and execution of serverless workflows using declarative YAML-based configuration files. This service supports both sequential and parallel execution patterns and integrates seamlessly with a wide range of Google Cloud services. Within the context of metadata enrichment pipeline, Google Cloud Workflows functions as the central orchestrator, coordinating the execution of multiple processing steps through services such as Cloud Functions and Eventarc. Its event-driven architecture allows workflows to be automatically initiated by data events originating from text databases, to make the metadata extraction streamlined from the batch of new documents.

*Batch Prediction with Gemini* [4] provides a cost-effective and efficient solution for large-scale language model inference tasks. This service is designed for high-throughput, non-latency-critical workloads, allowing users to submit a large number of prompts in a single asynchronous request. To utilize this capability, prompts are pre-constructed into a single file and the batch job is submitted, with results being retrieved once the processing is complete, typically within 24 hours.

We broadly divided the metadata enrichment process into two stages. In the first stage, we handle five metadata that rely on input text from individual files. This is followed by the second stage, where we process seven metadata that require input from multiple files. For each stage, we created several *Batch Prediction* jobs, each containing thousands of LLM inference requests. Specifically, each request corresponds to a metadata type and file, containing an optimized prompt and the required context text from the file. This process is repeated across all combinations of metadata and files to generate multiple requests. Typically, we package each batch job with $100,000$ to $120,000$ requests and create several batches. This orchestration approach enables scalable, fault-tolerant processing pipelines that can handle varying data volumes while maintaining consistent execution patterns and error handling capabilities across distributed cloud services.

## 3 Technical Validation

This section focuses on validating the accuracy and reliability of metadata enrichment methods. We assess the performance of these methods using expert-curated datasets to make sure they are extracting meaningful metadata.

### 3.1 Validating Metadata Enrichment Methods

We validate metadata enrichment methods, focusing on CE, EA and EIS documents.

#### *3.1.1 Validation on CE Documents*

The CE metadata enrichment methods are validated using an expert-curated dataset comprising 189 projects (166 from DOE and 23 from BLM) with annotations across multiple metadata categories, which serve as ground-truth labels for evaluation. For DOE and USDA projects, a single document is associated with each project, while BLM projects often consist of multiple files. In the latter case, selecting the correct candidate file is the most critical step, as errors in this stage propagate to metadata extraction. We therefore evaluate our candidate file selection process on a representative set of 23 BLM projects spanning all states included in the dataset.

Metadata quality is assessed by comparing the extracted metadata with the validation ground truth using the FuzzyWuzzy library[9] `token_set_ratio` metric. This metric computes similarity by tokenizing strings, removing duplicate words, and measuring overlap across tokens, making it robust to word order differences and extraneous terms. We consider a similarity score of 80% or higher to constitute a correct match, hereafter referred to as a *fuzzy match*. This metric is especially suitable for metadata fields with open-text values, while also being applicable to fixed categorical attributes.

To achieve high accuracy, we iteratively refined and optimized prompts until the validation set yielded reliable results. A prompt is deemed finalized once it consistently achieves more than 90% *fuzzy match* accuracy for the targeted metadata attribute.

Table 3 summarizes the performance on extracting CE specific metadata attributes. Each attribute presents unique challenges. For instance, finding the correct `project_type` in DOE CE documents can be difficult because the `action_description`

[4] https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/batch-prediction-gemini

| | Department of Energy | Bureau of Land Management |
|---|---|---|
| **Metadata Attribute** | **Accuracy (#correct/#total)** | **Accuracy (#correct/#total)** |
| project_title | 97.59% (162/166) | 95.24% (20/21) |
| project_type | 91.57% (76/83) | 95.24% (20/21) |
| project_sponsor | 95.12% (156/164) | 90.48% (19/21) |
| location | 97.57% (161/165) | 100.00% (21/21) |
| lead_agency | 98.79% (163/165) | 100.00% (21/21) |
| document_title | 95.78% (159/166) | 100.00% (21/21) |
| prepared_by | 93.98% (156/166) | 100.00% (21/21) |
| ce_category | 100.00% (153/153) | **100.00% (23/23)** |
| action_description | 100.00% (40/40) | **100.00% (22/22)** |

**Table 3.** The table reports the *fuzzy match* accuracy of our metadata extraction relative to the expert-annotated validation set. "**Accuracy (#correct/#total)**" corresponds to the percentage and count of correct extractions over the total number of ground-truth annotations.

| | Environmental Assessment | Environmental Impact Statement |
|---|---|---|
| **Metadata Attribute** | **Accuracy (#correct/#total)** | **Accuracy (#correct/#total)** |
| project_title | **100% (12/12)** | 91% (111/122) |
| project_type | 91.67% (11/12) | 90.83% (99/109) |
| project_sponsor | 91.67% (11/12) | 91.67% (110/120) |
| location | 91.67% (11/12) | 95.00% (112/118) |
| lead_agency | 91.67% (11/12) | 92.31% (96/104) |
| document_title | 91.67% (11/12) | 96.92% (63/65) |
| document_type | **100% (12/12)** | 96.92% (283/292) |
| prepared_by | 91.67% (11/12) | **100% (11/11)** |
| volume_or_section_title | 91.67% (11/12) | 90.70% (78/96) |
| main_document | **100% (12/12)** | 98.99% (98/99) |

**Table 4.** The table reports the *fuzzy match* accuracy of our metadata extraction relative to the expert-annotated validation set for EA and EIS. "**Accuracy (#correct/#total)**" corresponds to the percentage and count of correct extractions over the total number of ground-truth annotations.

often contains general terms that make it prone to falling into broad categories, such as *"Research and Development"* or *"Routine Maintenance"*. In some cases, the description implies concepts (e.g., nuclear-related activities) without explicitly mentioning key terms, requiring second-degree inference. Project descriptions may also map ambiguously to similar categories, such as *"Habitat Conservation Plan"* versus *"Ecosystem Management and Restoration"*. Through iterative prompt refinement guided by validation results, we improved `project_type` extraction accuracy from 48.19% with an initial prompt to 91.57% with a more sophisticated yet general prompt (see Section C.1 for final CE prompts). A similar refinement strategy was applied to other metadata attributes, ultimately yielding high accuracy across the validation set.

### 3.1.2 Validation on EA and EIS Documents

The validation of both EA and EIS metadata enrichment pipelines was also conducted using curated datasets of expert-annotated documents. The validation datasets comprised subject matter expert annotations across ten metadata categories, providing ground truth labels for systematic evaluation across EA and EIS documents. Due to the substantial cost and complexity of obtaining cross-annotated ground truth data from multiple subject matter experts (SME), the EA validation dataset was carefully curated to include only the most confident annotations, while the EIS validation benefited from a larger pool of ground truth annotations, though annotation quality varied across different metadata categories. Two primary evaluation metrics were employed: *exact match accuracy*, which required a perfect match between extracted and ground truth values, and *fuzzy match accuracy*, which allowed for 80% token similarity to accommodate natural variations in textual representation. This dual-metric approach was particularly important given the heterogeneous nature of environmental compliance metadata, where some fields represent fixed categorical values while others contain open-text descriptions that may exhibit semantic equivalence despite syntactic differences.

The validation process revealed significant performance improvements through systematic prompt refinement and enhanced text preprocessing, with valuable lessons learned from EIS prompt development being successfully transferred to EA extraction methodologies. The Prepared By metadata, which identifies the agency or organization responsible for preparing environmental documents, presented particular extraction challenges across both EA and EIS, as language models frequently confused it with the Project Sponsor who proposed or funded the project. Initial EA performance for this metadata showed only 25% exact match accuracy, but improved dramatically to 83.33% exact match accuracy following targeted prompt optimization that explicitly differentiated between document preparers and project sponsors—techniques refined through prior EIS development. Similarly, the Lead Agency metadata demonstrated consistent improvement from 41.67% to 91.67% accuracy in EA through iterative prompt development, with the final prompt incorporating detailed instructions for agency identification. This cross-pollination of prompt engineering insights enabled the EA pipeline to achieve high accuracy despite working with a more constrained ground truth dataset.

Beyond traditional accuracy metrics, the validation process incorporated specialized evaluation approaches tailored to specific metadata characteristics. For certain metadata fields like Project Location, defining an effective fuzzy match score threshold proved challenging due to the diverse ways geographic information could be expressed. Consequently, manual inspection by SME was employed as a more reliable quantification method for EIS location extraction, involving experts who assessed the correctness of LLM-extracted metadata by examining the expert-provided annotation, the extracted result with corresponding reasoning, and the source document. This manual inspection process, while resource-intensive, provided a more nuanced evaluation for complex geographic and categorical metadata where automated similarity measures proved insufficient. After evaluating multiple thresholds, an 80% token similarity was selected for fuzzy match accuracy, as it proved to be effective for extracting open-text metadata.

Different metadata categories exhibited distinct performance patterns reflecting their inherent characteristics and extraction complexity across both EA and EIS documents. The metadata fields were categorized into fixed categorical values (such as Project Type and Document Type), open-text fields (like Volume or Section Title and Prepared By), and binary classifications (such as Main Document requiring Yes/No responses). Fixed categorical metadata demonstrated equivalent exact and fuzzy match accuracies, with Project Type extraction benefiting from structured prompts that emphasized the hierarchical "Primary Category – Subcategory" classification system and specific instructions for identifying project scope and technical elements. Open-text metadata presented greater extraction challenges, with Volume or Section Title showing an initial accuracy of 0% exact match in EA that improved to 25% exact match and 91.67% fuzzy match accuracy through iterative prompt refinement, addressing common section identification failures. To further enhance performance, a *top-k* approach was implemented for Project Type extraction, where the language model generated multiple candidate answers instead of just one, providing an additional improvement beyond the earlier prompt optimizations, as the correct classification was typically present among the top candidates even when not ranked first.

The validation process revealed that extraction performance varied significantly with document page coverage, leading to the development of metadata-specific analysis strategies that balanced computational efficiency with extraction accuracy. Some metadata fields, such as Project Type and Project Location, showed improved extraction when analyzing more pages of both EA and EIS documents, capturing contextual information distributed throughout the document body. In contrast, metadata like Document Type and Document Title achieved good results with focused analysis of initial pages such as cover pages and headers, where this information is typically standardized and prominently displayed. The Document Type classification achieved good performance across EA and EIS documents by leveraging clear document structure indicators found on cover pages, document headers, and standardized terminology, with prompts designed to distinguish between different environmental review documents such as CE, EA and EIS.

This systematic approach to prompt engineering and validation-driven refinement successfully achieved the target of over 90% fuzzy match accuracy for all planned metadata extractions across both EA and EIS document types, as demonstrated in Table 4. The EA validation achieved perfect accuracy (100%) for Project Title, Document Type, and Main Document classification, while maintaining consistent performance above 91% for all other metadata categories. The EIS validation demonstrated complementary strengths, achieving perfect accuracy for Prepared By extraction and strong performance across all categories, with Location achieving 95% accuracy. The effectiveness of transferring prompt engineering insights from EIS to EA development enabled high-quality metadata extraction despite the more limited EA ground truth dataset, validating the systematic approach to environmental compliance document processing.

### 3.2 Subject Matter Experts Driven Validation for Extracted Metadata

In order to ensure that the quality of the metadata extracted for NEPATEC 2.0 is high, we performed a human validation of the extracted metadata by having Subject Matter Experts (SMEs) evaluate a sample of the extracted metadata.

### *3.2.1 Curating Validation Sample*

From the full NEPATECv2 document corpus, we sampled a total of 643 documents. These documents were selected to be representative of the overall population. The sampling was generally proportional to the overall population, but challenging cases for LLMs were weighted higher. These included CE documents with multiple CE codes and projects involving multiple documents. For CE (55k total documents), we sampled 30% DOE and 70% BLM. CE codes were sampled proportionally, with higher weight given to multi-code entries. Project types were sampled proportionally. For EIS (50k total documents), sampling was proportional to agency distribution and the number of documents per project. For EA (20k total documents), sampling followed the same procedure as EIS. Overall, we sampled 354 CE documents (210 BLM + 144 DOE), 109 EA, and 180 EIS documents for human evaluation.

### *3.2.2 Validation Procedure*

**Annotator Effort:** We identified 10 human evaluators, each assigned 12 hours of review. Each document contained 8–12 metadata fields, and evaluators referenced the source PDF to check correctness. The evaluation time averaged 1̃2 minutes per document.

**Logistics:** Each evaluator received a folder containing the PDF document and an Excel spreadsheet with one extracted metadata field per column. Evaluators entered their evaluation score in the adjacent column.

### *3.2.3 Scoring:*

Evaluators reviewed the extracted metadata against the source document and assigned a Human Evaluation Score on a 1–5 scale for each metadata field. Scores were defined as follows:

**Table 5.** The scoring system used by the human evalutors to judge the quality of NEPATEC 2.0 data.

| Score | Interpretation |
|---|---|
| 1 | Unusable: Mostly incorrect or missing, preventing downstream use. |
| 2 | Poor: Major errors or omissions; partial utility only. |
| 3 | Acceptable: Correct for key fields, with minor issues. |
| 4 | Good: Accurate and complete; slight formatting inconsistencies. |
| 5 | Excellent: Fully accurate, complete, and formatted ideally. |

### *3.2.4 Validation Results*

The validation results was overwhelmingly positive. Across all metadata fields and document types for our sample, the average mean score was approximately 4.48, the median score was 5.0, and the average standard deviation was about 1.02, indicating generally high accuracy with relatively low variability.

This high performance continues to hold when we look into individual categories. For all categories (BLM CEs, DOE CEs, EAs, and EISs), the human evaluation scores for metadata fields were consistently high, with mean scores generally between 4.3 and 4.9 out of 5. Median scores were typically 5, indicating that most fields were judged accurate and complete. Standard deviations were generally below 1.3, showing relatively low variability across evaluations.

BLM CEs and DOE CEs showed slightly more variation for certain attributes such as *project_type* and *ce_category*. For BLM CEs, most attributes scored well, but *ce_category* and *project_type* trailed slightly relative to others. For DOE CEs, scores were consistently high across attributes, with only minor variation in *project_type* and *location*.

EISs achieved the most consistently high scores across all fields. For EISs, results were the strongest and most uniform, with nearly all attributes averaging close to 5 and showing minimal variance, and even the lowest-scoring attributes, such as section_or_volume_title, averaging above 4.3.

EAs also performed strongly, with most attributes averaging above 4.4 and prepared_by reaching 4.0. The *prepared_by* field was a clear outlier with somewhat lower than others but still averaging around 4.0.

Overall, the evaluation indicates strong accuracy of extracted metadata, with only a few attributes showing modest room for improvement. The full breakdown of the results is shown in Tables 6, 7, 8, and 9.

## 3.3 Known Limitations

We identified various limitations related to prompting and extractions during metadata validations. These limitations should be considered when analyzing and interpreting the data and in the development of subsequent versions of the NEPATEC metadata.

**Table 6.** Summary statistics of the Human Evaluation Score for CE - BLM

| Metadata Attribute | Count | Mean Score | Median Score | Standard Deviation |
|---|---|---|---|---|
| document_type | 210 | 4.80 | 5 | 0.88 |
| project_title | 210 | 4.76 | 5 | 0.87 |
| project_sponsor | 209 | 4.63 | 5 | 1.11 |
| location | 209 | 4.56 | 5 | 1.01 |
| document_title | 210 | 4.41 | 5 | 1.20 |
| section_or_volume_title | 210 | 4.40 | 5 | 1.18 |
| prepared_by | 210 | 4.30 | 5 | 1.17 |
| lead_agency | 209 | 4.27 | 5 | 1.33 |
| action_description | 209 | 4.24 | 5 | 1.02 |
| project_type | 209 | 3.92 | 5 | 1.25 |
| ce_category | 210 | 3.76 | 4 | 1.26 |

**Table 7.** Summary statistics of the Human Evaluation Score for CE - DOE

| Metadata Attribute | Count | Mean Score | Median Score | Standard Deviation |
|---|---|---|---|---|
| document_type | 144 | 5.00 | 5 | 0.00 |
| lead_agency | 144 | 4.90 | 5 | 0.32 |
| project_title | 143 | 4.86 | 5 | 0.63 |
| document_title | 144 | 4.77 | 5 | 0.54 |
| prepared_by | 144 | 4.73 | 5 | 0.73 |
| action_description | 144 | 4.53 | 5 | 0.79 |
| ce_category | 144 | 4.42 | 5 | 1.09 |
| section_or_volume_title | 144 | 4.41 | 5 | 1.31 |
| location | 141 | 4.21 | 5 | 1.33 |
| project_sponsor | 144 | 4.03 | 5 | 1.54 |
| project_type | 141 | 3.80 | 4 | 1.31 |

**Table 8.** Summary statistics of the Human Evaluation Score for EA

| Metadata Attribute | count_eval | mean_score | median_score | std_score |
|---|---|---|---|---|
| project_sponsor | 108 | 4.78 | 5 | 0.87 |
| project_title | 109 | 4.67 | 5 | 0.99 |
| document_type | 109 | 4.65 | 5 | 1.12 |
| project_type | 109 | 4.62 | 5 | 0.99 |
| main_document | 108 | 4.48 | 5 | 1.35 |
| location | 109 | 4.48 | 5 | 1.06 |
| lead_agency | 109 | 4.43 | 5 | 1.31 |
| document_title | 109 | 4.36 | 5 | 1.37 |
| section_or_volume_title | 109 | 4.13 | 5 | 1.39 |
| prepared_by | 109 | 4.01 | 5 | 1.37 |

### 3.3.1 Data Downloading Issues

One notable limitation faced during the data collection process was the inability to download some documents from the sources. While we have access to URLs that are expected to point to downloadable PDF files, some URLs did not return the files. This issue likely arose due to factors, such as broken links, server errors, or a change in file location after the URLs were published (page not found). The proportion of non-downloadable links varies with the sources with roughly $3-4$ % overall.

### 3.3.2 Document Formats and PDF Parsing

Automated text extraction from PDF files presents a significant challenge, particularly when dealing with legacy NEPA documents. A primary obstacle in parsing these documents involves font encoding. A PDF file should contain a font map that

**Table 9.** Summary statistics of the Human Evaluation Score for EIS

| Metadata Attribute | count_eval | mean_score | median_score | std_score |
|---|---|---|---|---|
| project_sponsor | 178 | 4.96 | 5 | 0.25 |
| project_type | 179 | 4.94 | 5 | 0.33 |
| project_title | 179 | 4.89 | 5 | 0.52 |
| location | 179 | 4.81 | 5 | 0.58 |
| document_title | 179 | 4.68 | 5 | 0.95 |
| document_type | 179 | 4.63 | 5 | 0.87 |
| prepared_by | 179 | 4.51 | 5 | 1.04 |
| main_document | 179 | 4.48 | 5 | 1.20 |
| lead_agency | 179 | 4.46 | 5 | 1.19 |
| section_or_volume_title | 179 | 4.34 | 5 | 1.17 |

links the visual representation of a character to a standard character code, such as Unicode, which allows for accurate copying and pasting. However, in many legacy documents, this mapping can be absent, incomplete, or non-standard.

Furthermore, the nature of legacy government documents frequently involves scanned images of paper records. In these cases, Optical Character Recognition (OCR) is used to convert the image of the text into machine-readable text. This process often adds an invisible layer of text over the image.The accuracy of this OCR layer is highly dependent on the quality of the original scan; older, degraded, or poorly scanned documents can result in significant errors (see Figure 4).

Parsing CE documents presents several challenges. For example, CE documents from DOE frequently contain fillable fields, dropdown menus, and checkboxes, which standard PDF parsers often fail to extract reliably. Even when such fields are extracted, the output frequently lacks contextual information relative to the surrounding text. For example, in Figure 5, the fields `project_title`, `action_description`, `project_sponsor`, and `location` are present as fillable fields. CE codes may be represented as dropdown menus (Figure 5b) in some documents, while in others they are indicated using checkboxes (Figure 5a).

Preserving the original document layout is therefore critical, as CE codes can appear in multiple formats. Apart from being embedded in checkboxes or dropdowns, they may also be presented as plain text. For instance, in Figure 6a, the CE code is shown on the left, with its corresponding description on the right.

Another challenge arises when text is embedded within images. This may occur across the entire PDF or within specific sections (see Figure 6a, where the header contains the document title inside an image). In such cases, OCR-based text extraction is necessary to recover the information.

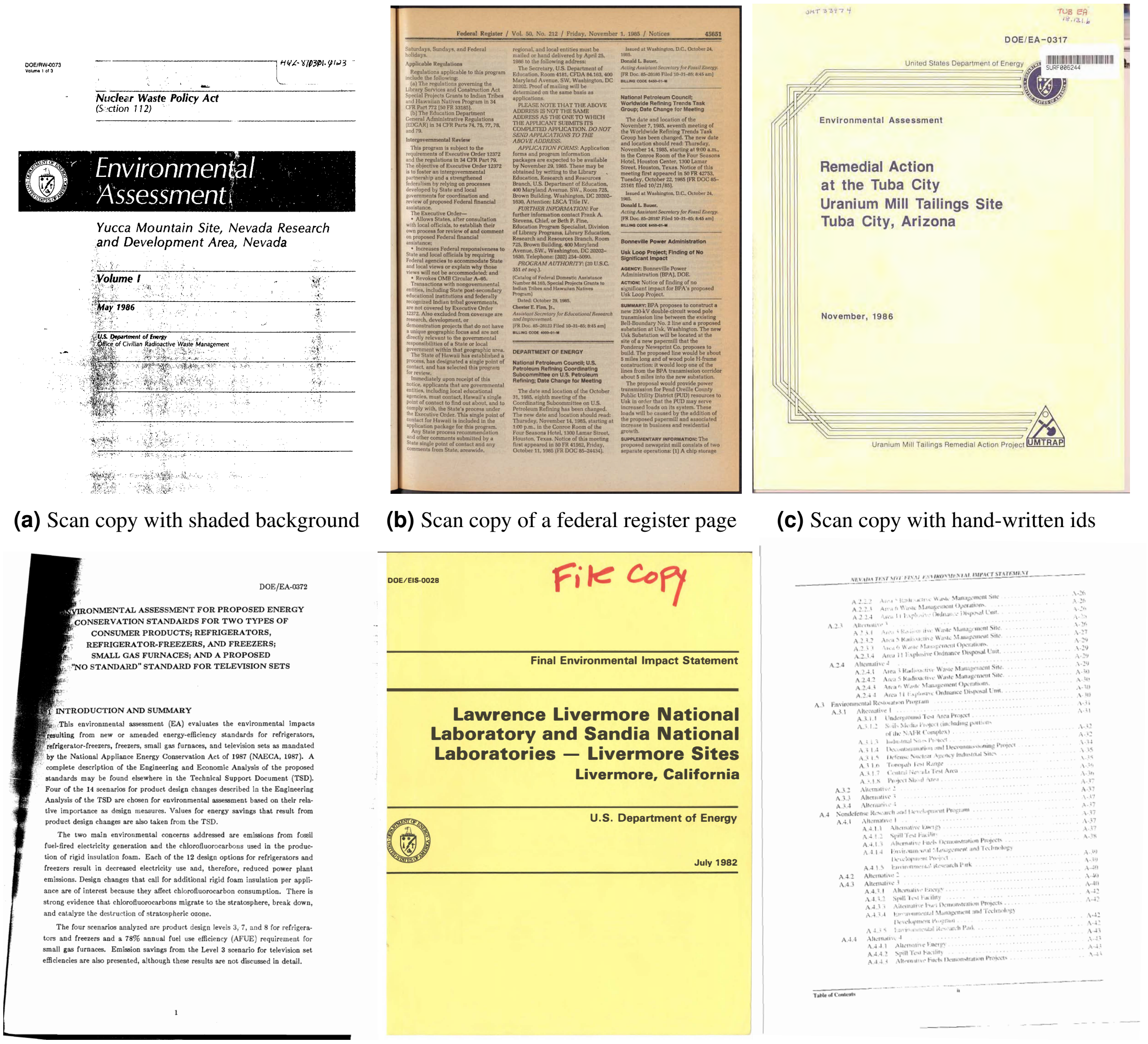
However, OCR alone is not always sufficient, as CE documents often exhibit significant variation in layout. Figure 6 demonstrates how the overall structure can differ considerably from one document to another. For example, the document title may appear embedded within an image or, in some cases, at the top of the page as part of the header. We noted that preserving the relative positioning of text elements is important for enabling accurate metadata extraction.

To address these challenges, we employ the Gemini-2.5 multi-modal model[7] in combination with prompt-based methods that integrate OCR, PDF parsing, and LLMs. This approach allows us to retain contextual information about fillable fields, extract text embedded within images, and maintain document formatting more faithfully.

#### 3.3.3 Duplicate copies of NEPA documents exist in different government websites

It can be challenging to recognize and properly reconcile multiple copies of a single NEPA document received from different government sources. This situation may arise when a lead agency files an EIS with the EPA but also maintains a copy in its own agency database. An agency may also adopt a NEPA document prepared by another federal agency and maintain its own identical or near-identical copy. The Thacker Pass project, a proposed lithium mine in Humboldt County, Nevada, provides a real-world illustration of challenges associated with LLM interpretations of duplicate NEPA documents in the case of an interagency adoption. The BLM, as the agency responsible for managing the public lands where the proposed mine would be located, was the lead agency for the initial NEPA review of the applicant's mine and exploration plans. In December 2020, the BLM issued a final FEIS for the project. The DOE adopted the BLM's analysis for its independent purposes of considering the provision of a loan to Lithium Nevada Corp. for the construction of the mine's processing facilities.

Despite the efficiencies of the NEPA adoption process, the Thacker Pass case also sheds light on the challenges of integrating



DOE/RW-0073
Volume 1 of 3

HQZ·870301·9123

Nuclear Waste Policy Act
(Section 112)

Environmental Assessment

Yucca Mountain Site, Nevada Research and Development Area, Nevada

Volume I

May 1986

U.S. Department of Energy
Office of Civilian Radioactive Waste Management

**(a)** Scan copy with shaded background **(b)** Scan copy of a federal register page

DOE/EA-0317

United States Department of Energy

SURF 005244

Environmental Assessment

Remedial Action at the Tuba City Uranium Mill Tailings Site Tuba City, Arizona

November, 1986

Uranium Mill Tailings Remedial Action Project

UMTRAP

**(c)** Scan copy with hand-written ids

DOE/EA-0372

ENVIRONMENTAL ASSESSMENT FOR PROPOSED ENERGY CONSERVATION STANDARDS FOR TWO TYPES OF CONSUMER PRODUCTS; REFRIGERATORS, REFRIGERATOR-FREEZERS, AND FREEZERS; SMALL GAS FURNACES; AND A PROPOSED "NO STANDARD" STANDARD FOR TELEVISION SETS

1 INTRODUCTION AND SUMMARY

This environmental assessment (EA) evaluates the environmental impacts resulting from new or amended energy-efficiency standards for refrigerators, refrigerator-freezers, freezers, small gas furnaces, and television sets as mandated by the National Appliance Energy Conservation Act of 1987 (NAECA, 1987). A complete description of the Engineering and Economic Analysis of the proposed standards may be found elsewhere in the Technical Support Document (TSD). Four of the 14 scenarios for product design changes described in the Engineering Analysis of the TSD are chosen for environmental assessment based on their relative importance as design measures. Values for energy savings that result from product design changes are also taken from the TSD.

The two main environmental concerns addressed are emissions from fossil fuel-fired electricity generation and the chlorofluorocarbons used in the production of rigid insulation foam. Each of the 12 design options for refrigerators and freezers result in decreased electricity use and, therefore, reduced power plant emissions. Design changes that call for additional rigid foam insulation per appliance are of interest because they affect chlorofluorocarbon consumption. There is strong evidence that chlorofluorocarbons migrate to the stratosphere, break down, and catalyze the destruction of stratospheric ozone.

The four scenarios analyzed are product design levels 3, 7, and 8 for refrigerators and freezers and a 78% annual fuel use efficiency (AFUE) requirement for small gas furnaces. Emission savings from the Level 3 scenario for television set efficiencies are also presented, although these results are not discussed in detail.

1

DOE/EIS-0028

File Copy

Final Environmental Impact Statement

Lawrence Livermore National Laboratory and Sandia National Laboratories — Livermore Sites
Livermore, California

U.S. Department of Energy

July 1982

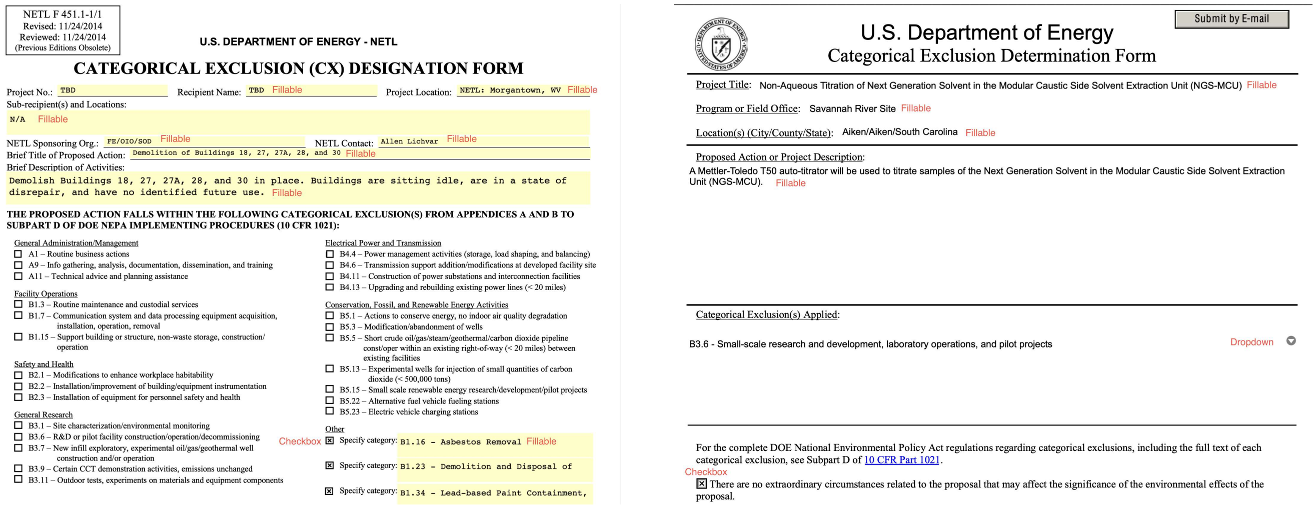
**(d)** Scan pages with dark background **(e)** Scan file copy **(f)** Scan table of content

**Figure 4.** Various scan formats of NEPA documents

documents gathered across multiple government websites. When several agencies—in this instance, the DOE, and BLM [5]—are involved in a single project, each may host project-related documents on their respective websites (see Figure 7), in addition to the copies maintained in the EPA's EIS database [6]. In this case, because DOE did not participate as a cooperating agency in the preparation of the EIS, DOE re-circulated a copy of the final BLM EIS with several added cover pages as a final DOE EIS (DOE/EIS-0561) [7] for a period of 30 days, then also filed a copy of the final DOE EIS with the EPA. The added cover pages identify DOE as the lead federal agency on page 2 of the document for the purposes of its loan decision, which precede a copy of the final BLM EIS that, on page 7, states that the lead agency is the "U.S. Department of the Interior Bureau of Land Management Humboldt River Field Office." While technically both statements are correct in their respective contexts—that the BLM led the original EIS and the DOE is the lead agency for its adoption of the EIS—these statements can easily mislead an LLM that is not well-versed in the nuances of NEPA's interagency procedures.

In ongoing research, we will continue to explore techniques to detect and interpret relationships between duplicate or near-duplicate documents, determining whether they should be replaced with a single copy or labeled as an adoption. In cases

[5] https://eplanning.blm.gov/eplanning-ui/project/1503166/510
[6] https://cdxapps.epa.gov/cdx-enepa-II/public/action/eis/details?eisId=315942
[7] https://www.energy.gov/lpo/articles/eis-0561-final-environmental-impact-statement


Fillable
Checkbox
Dropdown

NETL F 451.1-1/1
Revised: 11/24/2014
Reviewed: 11/24/2014
(Previous Editions Obsolete)

U.S. DEPARTMENT OF ENERGY - NETL

CATEGORICAL EXCLUSION (CX) DESIGNATION FORM

Project No.: TBD   Recipient Name: TBD   Project Location: NETL: Morgantown, WV
Sub-recipient(s) and Locations:
N/A
NETL Sponsoring Org.: FE/OIO/SOD   NETL Contact: Allen Lichvar
Brief Title of Proposed Action: Demolition of Buildings 18, 27, 27A, 28, and 30
Brief Description of Activities:
Demolish Buildings 18, 27, 27A, 28, and 30 in place. Buildings are sitting idle, are in a state of disrepair, and have no identified future use.

THE PROPOSED ACTION FALLS WITHIN THE FOLLOWING CATEGORICAL EXCLUSION(S) FROM APPENDICES A AND B TO SUBPART D OF DOE NEPA IMPLEMENTING PROCEDURES (10 CFR 1021):

General Administration/Management
- □ A1 – Routine business actions
- □ A9 – Info gathering, analysis, documentation, dissemination, and training
- □ A11 – Technical advice and planning assistance

Facility Operations
- □ B1.3 – Routine maintenance and custodial services
- □ B1.7 – Communication system and data processing equipment acquisition, installation, operation, removal
- □ B1.15 – Support building or structure, non-waste storage, construction/ operation

Safety and Health
- □ B2.1 – Modifications to enhance workplace habitability
- □ B2.2 – Installation/improvement of building/equipment instrumentation
- □ B2.3 – Installation of equipment for personnel safety and health

General Research
- □ B3.1 – Site characterization/environmental monitoring
- □ B3.6 – R&D or pilot facility construction/operation/decommissioning
- □ B3.7 – New infill exploratory, experimental oil/gas/geothermal well construction and/or operation
- □ B3.9 – Certain CCT demonstration activities, emissions unchanged
- □ B3.11 – Outdoor tests, experiments on materials and equipment components

Electrical Power and Transmission
- □ B4.4 – Power management activities (storage, load shaping, and balancing)
- □ B4.6 – Transmission support addition/modifications at developed facility site
- □ B4.11 – Construction of power substations and interconnection facilities
- □ B4.13 – Upgrading and rebuilding existing power lines (< 20 miles)

Conservation, Fossil, and Renewable Energy Activities
- □ B5.1 – Actions to conserve energy, no indoor air quality degradation
- □ B5.3 – Modification/abandonment of wells
- □ B5.5 – Short crude oil/gas/steam/geothermal/carbon dioxide pipeline const/oper within an existing right-of-way (< 20 miles) between existing facilities
- □ B5.13 – Experimental wells for injection of small quantities of carbon dioxide (< 500,000 tons)
- □ B5.15 – Small scale renewable energy research/development/pilot projects
- □ B5.22 – Alternative fuel vehicle fueling stations
- □ B5.23 – Electric vehicle charging stations

Other
- ☒ Specify category: B1.16 - Asbestos Removal
- ☒ Specify category: B1.23 - Demolition and Disposal of
- ☒ Specify category: B1.34 - Lead-based Paint Containment,

**(a)** CE document with fillable fields.

Submit by E-mail

U.S. Department of Energy
Categorical Exclusion Determination Form

Project Title: Non-Aqueous Titration of Next Generation Solvent in the Modular Caustic Side Solvent Extraction Unit (NGS-MCU)

Program or Field Office: Savannah River Site

Location(s) (City/County/State): Aiken/Aiken/South Carolina

Proposed Action or Project Description:
A Mettler-Toledo T50 auto-titrator will be used to titrate samples of the Next Generation Solvent in the Modular Caustic Side Solvent Extraction Unit (NGS-MCU).

Categorical Exclusion(s) Applied:

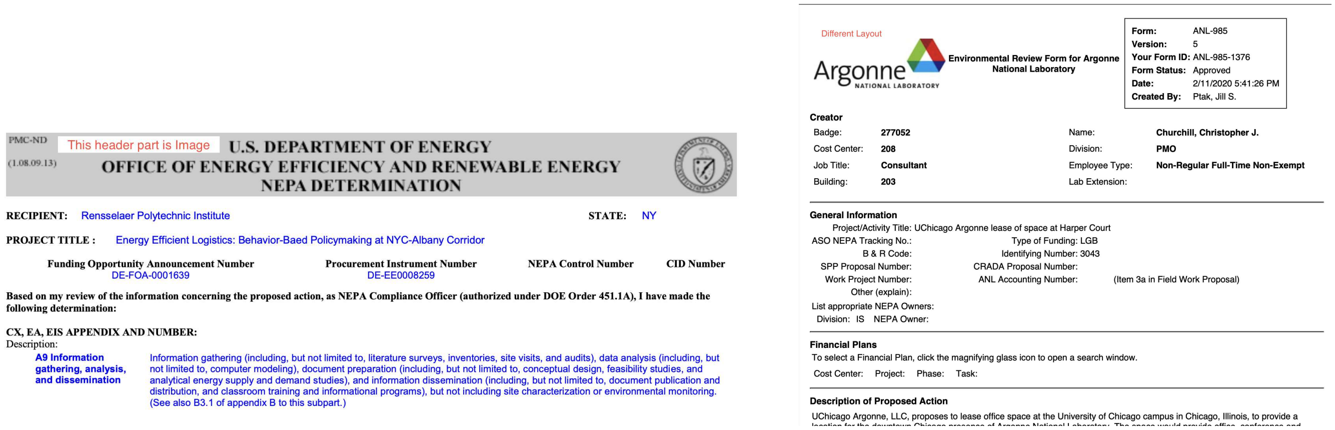
B3.6 - Small-scale research and development, laboratory operations, and pilot projects

For the complete DOE National Environmental Policy Act regulations regarding categorical exclusions, including the full text of each categorical exclusion, see Subpart D of 10 CFR Part 1021.

☒ There are no extraordinary circumstances related to the proposal that may affect the significance of the environmental effects of the proposal.

**(b)** CE document with ce_category as dropdown.

**Figure 5.** The figures present two DOE CE documents containing fillable fields, checkboxes, and dropdown menus.


This header part is Image
Different Layout

PMC-ND
(1.08.09.13)

U.S. DEPARTMENT OF ENERGY
OFFICE OF ENERGY EFFICIENCY AND RENEWABLE ENERGY
NEPA DETERMINATION

RECIPIENT: Rensselaer Polytechnic Institute   STATE: NY

PROJECT TITLE : Energy Efficient Logistics: Behavior-Baed Policymaking at NYC-Albany Corridor

| Funding Opportunity Announcement Number | Procurement Instrument Number | NEPA Control Number | CID Number |
|---|---|---|---|
| DE-FOA-0001639 | DE-EE0008259 | | |

Based on my review of the information concerning the proposed action, as NEPA Compliance Officer (authorized under DOE Order 451.1A), I have made the following determination:

CX, EA, EIS APPENDIX AND NUMBER:
Description:

**A9 Information gathering, analysis, and dissemination** Information gathering (including, but not limited to, literature surveys, inventories, site visits, and audits), data analysis (including, but not limited to, computer modeling), document preparation (including, but not limited to, conceptual design, feasibility studies, and analytical energy supply and demand studies), and information dissemination (including, but not limited to, document publication and distribution, and classroom training and informational programs), but not including site characterization or environmental monitoring. (See also B3.1 of appendix B to this subpart.)

**(a)** CE document with text inside the header as an image.

Argonne NATIONAL LABORATORY

Environmental Review Form for Argonne National Laboratory

Form: ANL-985
Version: 5
Your Form ID: ANL-985-1376
Form Status: Approved
Date: 2/11/2020 5:41:26 PM
Created By: Ptak, Jill S.

Creator

| | | | |
|---|---|---|---|
| Badge: | 277052 | Name: | Churchill, Christopher J. |
| Cost Center: | 208 | Division: | PMO |
| Job Title: | Consultant | Employee Type: | Non-Regular Full-Time Non-Exempt |
| Building: | 203 | Lab Extension: | |

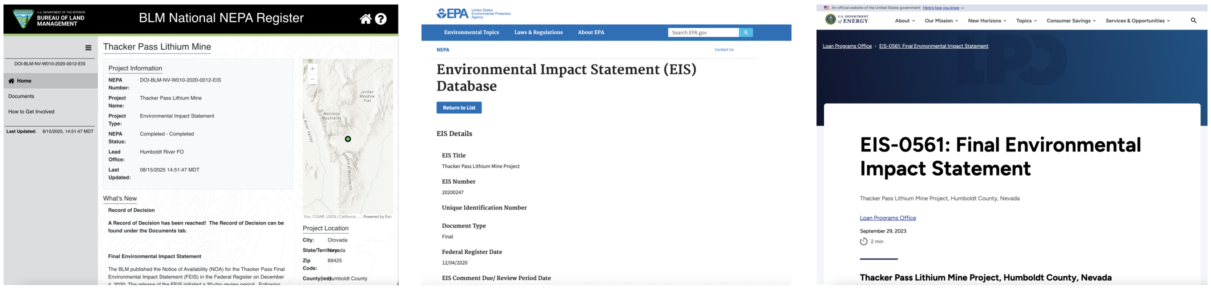
General Information
Project/Activity Title: UChicago Argonne lease of space at Harper Court
ASO NEPA Tracking No.: Type of Funding: LGB
B & R Code: Identifying Number: 3043
SPP Proposal Number: CRADA Proposal Number:
Work Project Number: ANL Accounting Number: (Item 3a in Field Work Proposal)
Other (explain):
List appropriate NEPA Owners:
Division: IS NEPA Owner:

Financial Plans
To select a Financial Plan, click the magnifying glass icon to open a search window.
Cost Center: Project: Phase: Task:

Description of Proposed Action
UChicago Argonne, LLC, proposes to lease office space at the University of Chicago campus in Chicago, Illinois, to provide a

**(b)** CE document with different layout.

**Figure 6.** The figures present two DOE CE documents containing different document layouts.


U.S. DEPARTMENT OF THE INTERIOR
BUREAU OF LAND MANAGEMENT
BLM National NEPA Register
Home
Documents
How to Get Involved
Jordan Meadow Flat
Montana Mountains
River Valley
Double H Mountains
Esri, CGIAR, USGS | California... Powered by Esri
EPA United States Environmental Protection Agency
Environmental Topics
Laws & Regulations
About EPA
Search EPA.gov
NEPA
Contact Us
Return to List
An official website of the United States government Here's how you know
U.S. DEPARTMENT of ENERGY
About
Our Mission
New Horizons
Topics
Consumer Savings
Services & Opportunities
Loan Programs Office › EIS-0561: Final Environmental Impact Statement

Thacker Pass Lithium Mine

DOI-BLM-NV-W010-2020-0012-EIS

Last Updated: 8/15/2025, 14:51:47 MDT

Project Information

| | |
|---|---|
| NEPA Number: | DOI-BLM-NV-W010-2020-0012-EIS |
| Project Name: | Thacker Pass Lithium Mine |
| Project Type: | Environmental Impact Statement |
| NEPA Status: | Completed - Completed |
| Lead Office: | Humboldt River FO |
| Last Updated: | 08/15/2025 14:51:47 MDT |

What's New

Record of Decision

A Record of Decision has been reached! The Record of Decision can be found under the Documents tab.

Final Environmental Impact Statement

The BLM published the Notice of Availability (NOA) for the Thacker Pass Final Environmental Impact Statement (FEIS) in the Federal Register on December 4, 2020. The release of the FEIS initiated a 30-day review period. Following

Project Location

| | |
|---|---|
| City: | Orovada |
| State/Territory: | Nevada |
| Zip Code: | 89425 |
| County(ies): | Humboldt County |

**(a)** BLM's Records

Environmental Impact Statement (EIS) Database

EIS Details

EIS Title
Thacker Pass Lithium Mine Project

EIS Number
20200247

Unique Identification Number

Document Type
Final

Federal Register Date
12/04/2020

EIS Comment Due/ Review Period Date

**(b)** EPA's Records

EIS-0561: Final Environmental Impact Statement

Thacker Pass Lithium Mine Project, Humboldt County, Nevada

Loan Programs Office

September 29, 2023
2 min

Thacker Pass Lithium Mine Project, Humboldt County, Nevada

**(c)** DOE's Records

**Figure 7.** Multiple Copies of Final Environmental Impact Statement, Thacker Pass Lithium Mine Project maintained by BLM, EPA and DOE

of interagency adoptions, we anticipate utilizing a parent process ID attribute to preserve agency specific information while associating the adopted files with one another.

#### 3.3.4 Existence of main NEPA documents within a project

We checked if each grouping of files for a given project included the main NEPA document, such as a CE, EA, or EIS. In the EA batch, projects had at least one of these records: *EA*, *DEA*, or *FONSI*. However, in the EIS batch, 25% of the projects did not have have any of the key records: *FEIS*, *DEIS*, or *ROD*. This might be due to mistakes in document classification or , more commonly, represents project for which the only available files are comment letters from the EPA's EIS database.

#### 3.3.5 Additional post-processing required for some metadata

Most extracted values are copied directly from the source text. Additional post-processing of the raw extracted values could further standardize the data and merge semantically equivalent but syntactically different values. For example, semantically equivalent variants are common in the *lead_agency* and *ce_category* values of some document types, but could be standardized to a constrained list of values. Other properties, such as *prepared_by* are not suitable for standardization due to the unbounded number of potential preparers, which include consulting firms and state agencies spanning multiple decades. *Action_description*, which is generally a raw extraction from the source CE document, could be further refined by imposing a maximum character limit and prompting an LLM to reform the raw text into a summary.

#### 3.3.6 Over-assignment of process types

*Process_type* and, the related *process_family* attribute by extension, are in some cases over-assigned relative to the intended behavior, which was to return only those project types that represent defining characteristics of the action. Some assigned types are unrelated or minor components, or related to potential impacts of the action rather than the action itself.

#### 3.3.7 Variable location identification methodologies

For better performance, we used different prompts to extract the *project_location* across CE (see Section C.1) and EA/EIS (see Section C.2 and C.3). Generally, when a project is described as being situated in one or more counties or states, the location attribute identifies each geographic area and provides the centroid of their combined geographic coverage (expressed in decimal degrees). Extracted location values from EA and EIS documents provide an inferred latitude and longitude, whereas location values from CE documents usually consist of text describing the action's location copied verbatim from the document. The EA and EIS extracted values have the advantage of being "map-ready," but do not capture more detailed locational descriptions such as Public Land Survey System descriptions (e.g., township, section, range) or types of areas other than city, county, or state (e.g., a watershed or agency administrative district). Improving the consistency of responses across document types and further consideration of the relative advantages of each approach is warranted.

#### 3.3.8 Contextual errors

For certain metadata attributes, there is no single correct solution on which to evaluate the extractions. For example, although simple in concept, *document_title* may be entirely absent, it may be unclear what specific text string the preparers intended to include in the title due to the formatting and placement of words on the cover page, or there may be inconsistencies in the document title between elements such as the cover page, title page, and headings. In such cases, the LLM must make a subjective inference or selection.

In some cases, extractions appear to indicate misinterpreted context or incorrect inferences. For example, values returned for the *project_sponsor* and *prepared_by* attributes may be present in the document text, but are neither sponsors nor preparers.

#### 3.3.9 Limited ranges of process and document types

In NEPATEC2.0, process type assignments are limited to CE, EA, and EIS. Future versions of the metadata may benefit from expansion to include variations such as adoption of another agency NEPA review and programmatic and supplemental NEPA documents. Similarly, the resolution of the data may be improved through further classification of documents now labeled as "Other" into distinct types, such as technical support documents, public engagement materials, or applications from project sponsors.

## 4 Usage

Use of LLMs offers myriad opportunities to gain valuable insights through reading, interpretation, and analysis of the NEPA documents in NEPATEC2.0, 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.

**Enhanced search and retrieval of NEPA documents.** The NEPATEC2.0 metadata enables rapid search of discovery of NEPA documents prepared by many Federal agencies based on fundamental characteristics such as lead agency, project type, project location, and process type. This makes it easier for agencies to reuse and adapt language and information compiled for previous NEPA reviews based on specific search characteristics. Metadata also enables high-level statistics on the quantity and characteristics of NEPA reviews, while the inclusion of page-wise text enables deeper analysis, such as identifying circumstances where air quality modeling was performed for NEPA analysis.

**Augmenting CEQ's CE Explorer Tool** CEQ's CE Explorer Tool[8] allows the public to perform a text search across 2,290 categories of actions that at least one federal agency has determined do not normally significantly affect the quality of the human environment. NEPATEC2.0 provides provides an opportunity to augment the search tool by also returning examples of completed CEs that apply select categories of interest or have similar project descriptions. NEPATEC2.0 contains more than **50,00** number of CE decisions made by DOE, BLM, and USDA. The repository of CEs in NEPATEC2.0 could also serve as points of reference for LLM agents tasked with drafting sections of future CE documents, such as project descriptions, documentation of compliance with NEPA, and analysis of extraordinary circumstances.

**Analysis of geographic distribution and project types.** NEPATEC2.0 provides a unique dataset for analyzing the distribution of NEPA reviews by geography and project type as shown in Figures 12 and 13. This information can be used to identify past NEPA reviews in areas considered for new actions, which may offer insights about environmental issues of concern, baseline conditions, locally relevant literature and reports, and impacts from past actions.

**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 provide an annotated list of the primary types of alternatives considered for oil and gas projects in Wyoming, or generating a list of potential mitigation measures for energy projects that may affect greater sage-grouse.

**Identifying scientific concepts.** 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. Exploration into 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, could help identify scientific studies germane to future reviews and highlight any gaps in existing research.

[8] `https://ce.permitting.innovation.gov/`

## 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. We would also like to thank Jordan Eccles, Pranava Raparla, Taresa Lawrence, Jack Titus, Keith Benes, Davie Nguyen, Stanley Mcmillian and all others who contributed to the discussion in developing NEPATECv2 metadata. We also want to thank Leah Hare, Lin Zeng, Caitlin Wessel, Caitlin Condon, Jim Jackson, Dave Goodman, Lindsey Renaud, and Cleve Davis who provided annotations validating a sample of NEPATECv2 metadata. We also want to thank Lav Jain, Jeffrey Ouimette, Dave Stone, Steven Boesel, and Beth Hartman for their help on deploying the NEPA metadata extraction framework in Google Cloud. This technical report has been cleared by PNNL for public release as PNNL-38163.

## References

1. CEQ. Length of environmental impact statements (2013-2018). https://ceq.doe.gov/docs/nepa-practice/CEQ_EIS_Length_Report_2020-6-12.pdf (2024). [Online; accessed 21-June-2021].

2. CEQ. Ceq permitting technology action plan. https://permitting.innovation.gov/CEQ_Permitting_Technology_Action_Plan.pdf (2025). [Online; accessed 10-Aug-2025].

3. Laparra, E. *et al.* Addressing structural hurdles for metadata extraction from environmental impact statements. *J. Assoc. for Inf. Sci. Technol.* **74**, 1124–1139 (2023).

4. Bethard, S. *et al.* 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).

5. CEQ. Council on environmental quality report to congress on the potential for online and digital technologies to address delays in reviews and improve public accessibility and transparency under 42 u.s.c. 4332(2)(c). https://ceq.doe.gov/docs/ceq-reports/CEQ-E-NEPA-Report-to-Congress_Final-(508).pdf (2025). [Online; accessed 10-Aug-2025].

6. CEQ. Ceq nepa and permitting data and technology standard. https://permitting.innovation.gov/CEQ_NEPA_and_Permitting_Data_and_Technology_Standard.pdf (2025). [Online; accessed 10-Aug-2025].

7. Comanici, G. *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

8. Khattab, O. *et al.* Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024* (2022).

9. SeatGeek. FuzzyWuzzy: Fuzzy String Matching in Python. https://github.com/seatgeek/fuzzywuzzy (2013–).

# A Data Records

## A.1 Data Repository and Access

The NEPATECv2 Dataset is publicly accessible at `https://huggingface.co/datasets/PNNL/NEPATEC2.0`.

## A.2 Repository Structure

Please see the data structure below for the JSON files available in the public repository. At the top level, it contains three main sections: **project, process, and documents**. The project section holds metadata about the project itself, such as ID, title, sector, type, description, sponsor, and location, each as an object with a value field. The process section describes the process family, type, and lead agency, also using value fields. The documents array contains multiple document objects, each with nested metadata (including both document and file metadata) and a pages array that contains the page texts. Note that, while all document objects have file metadata, they will not have the document metadata. For example, when an FEIS document is divided into volumes, we will record the document metadata only for the main file as we identified in the metadata extraction process. Each entry in the pages array can include either just page data (with page numbers and text) or additional metadata about the file, followed by page 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": ""
                            },
                            "file_provider": {
                                "value": ""
                            }
                        }
            },
            "pages": [
                {
                    "page number": 1,
                    "page text": "PAGE 1 TEXT"
                },
                {
                    "page number": 2,
                    "page text": "PAGE 2 TEXT"
                }
            ]
        },
        {
            "metadata": {
                "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": ""
                            },
                            "file_provider": {
                                "value": ""
                            }
                        }
            },
            "pages": [
                {
                    "page number": 1,
                    "page text": "PAGE 1 TEXT"
                },
                {
                    "page number": 2,
                    "page text": "PAGE 2 TEXT"
                }
            ]
        }
    ]
}
```

## B Appendix: Data Statistics and Insights

For research purposes, we performed multiple visualizations across various metadata to better understand the characteristics and patterns within the NEPATECv2 dataset. These visualizations include:

- **Distributions of project types:** Figure 8, Figure 9 and Figure 10 present the Top-20 most frequent project types in each process types of CE, EA and EIS, respectively.
- **Lead Agency involvement:** Figure 11 highlights the distribution of EIS projects across the Top-20 most frequent lead agencies, derived from documents obtained from the EPA.
- **Geographic Distribution of NEPA projects:** Figure 12 shows the geographic locations of EA projects across the United States, using markers to represent project centroids at the county level. Similarly, Figure 13 provides a visualization of the spatial distribution of EIS projects, showcasing project clustering and regional variation.
- **Distribution of CE categories:** Figure 14 analyzes the distribution of CE projects across different CE categories, specifically within DOE documents from the NEPATECv2 dataset.

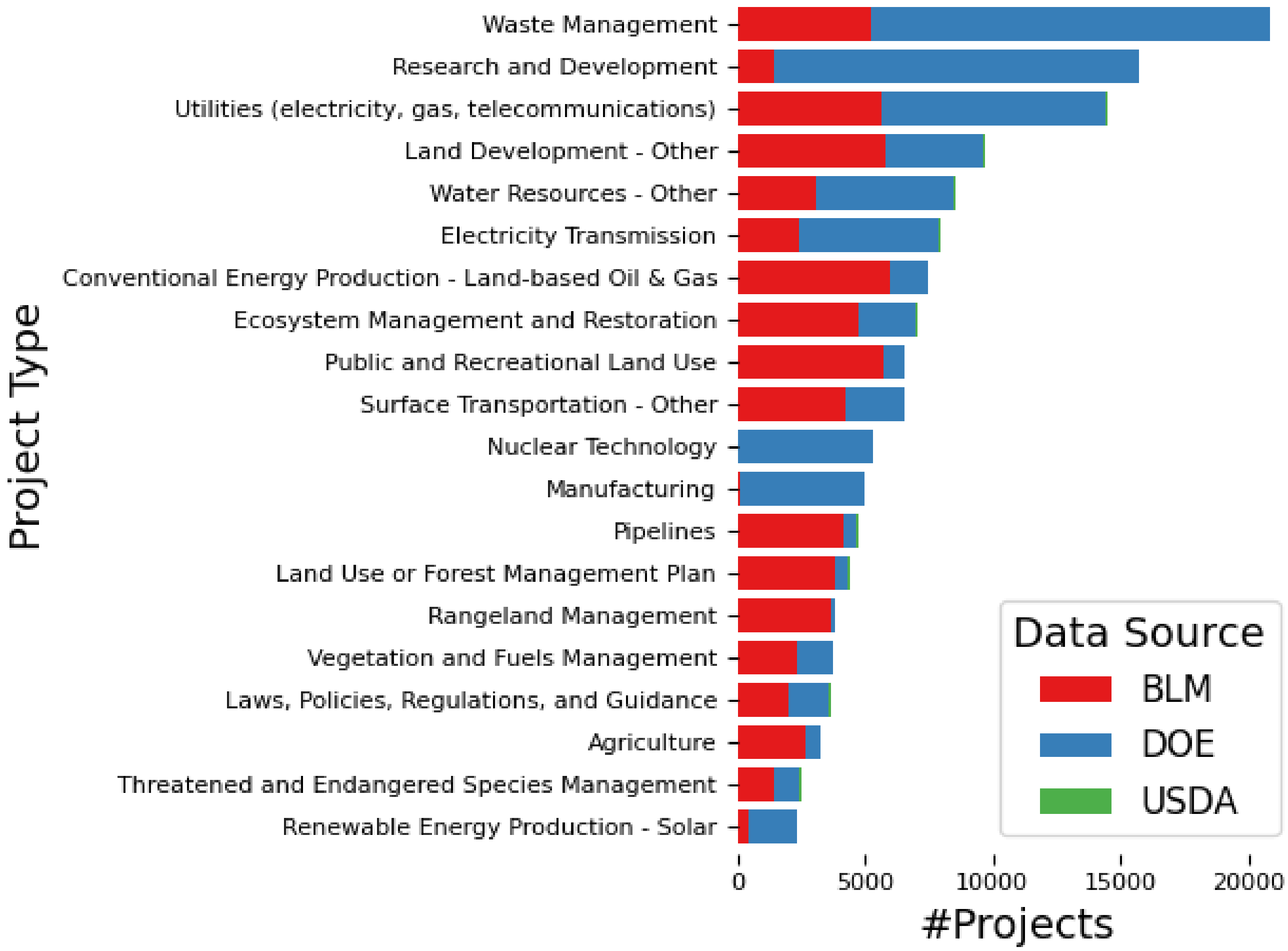


**Figure 8.** Distribution of CE projects across project types. For brevity, we only visualize the Top-20 most frequent project types in CE documents from the NEPATECv2 dataset.

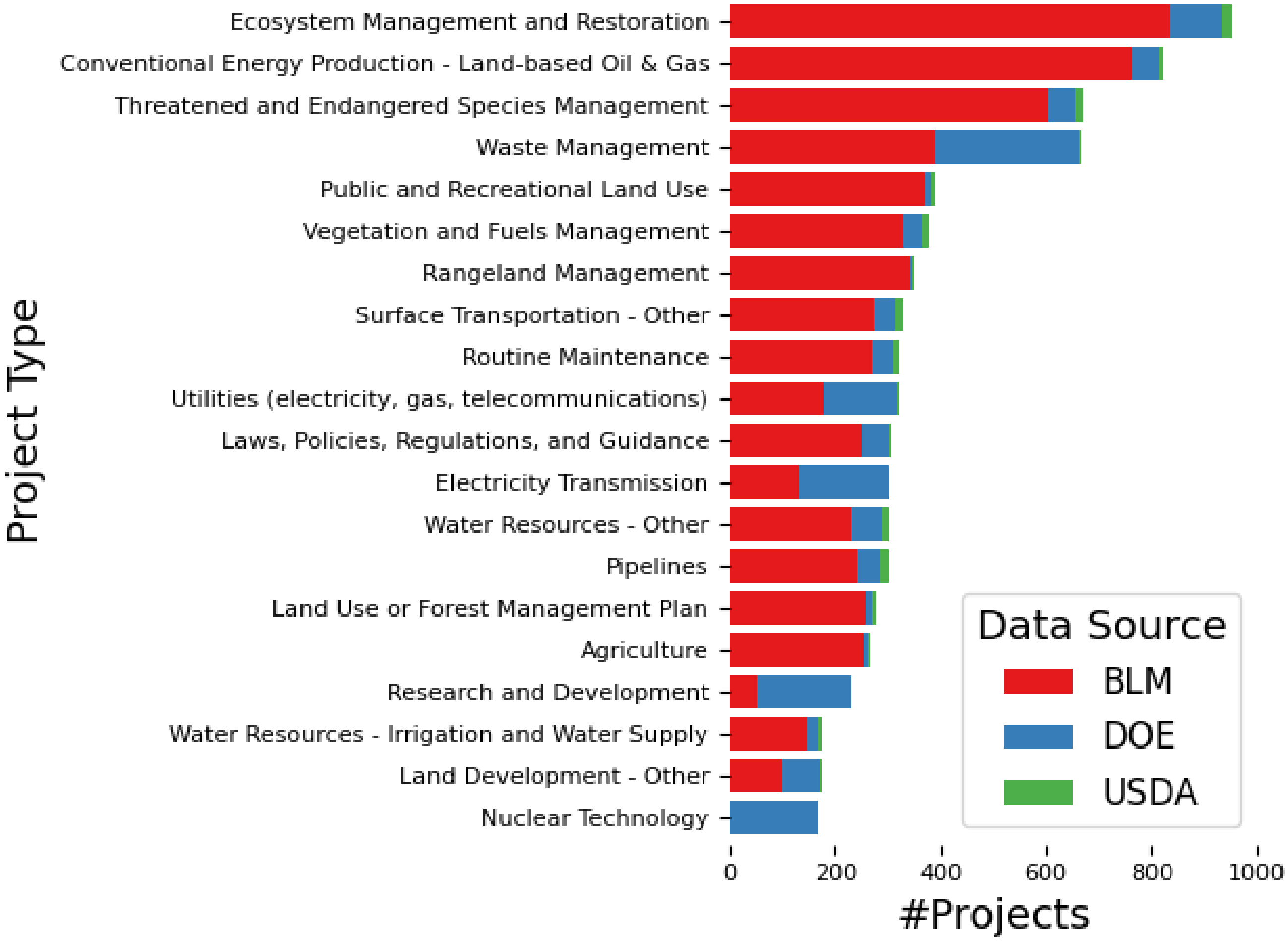


**Figure 9.** Distribution of EA projects across project types. For brevity, we only visualize the Top-20 most frequent project types in EA documents from the NEPATECv2 dataset.

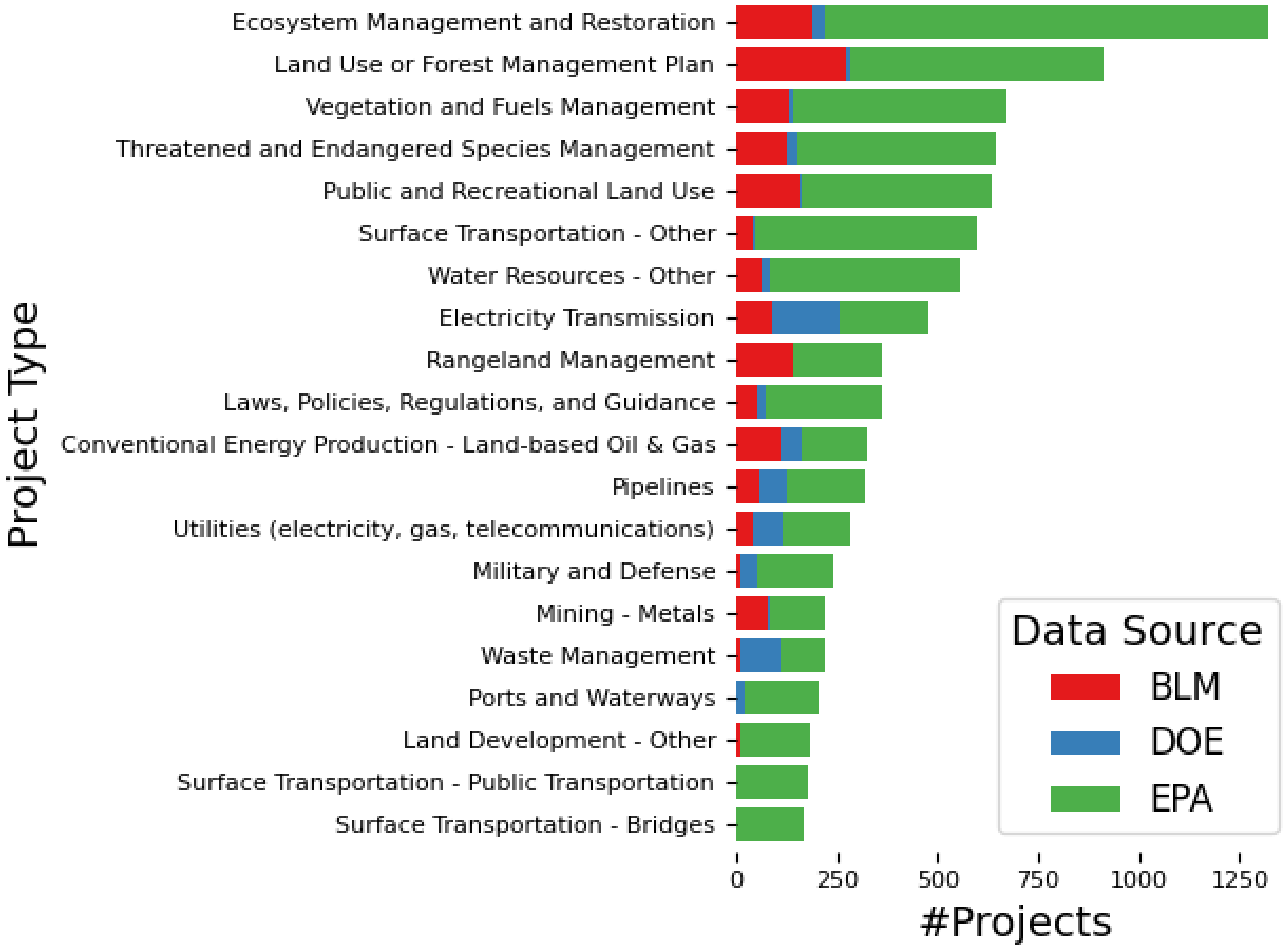


**Figure 10.** Distribution of EIS projects across project types. For brevity, we only visualize the Top-20 most frequent project types in EIS documents from the NEPATECv2 dataset.

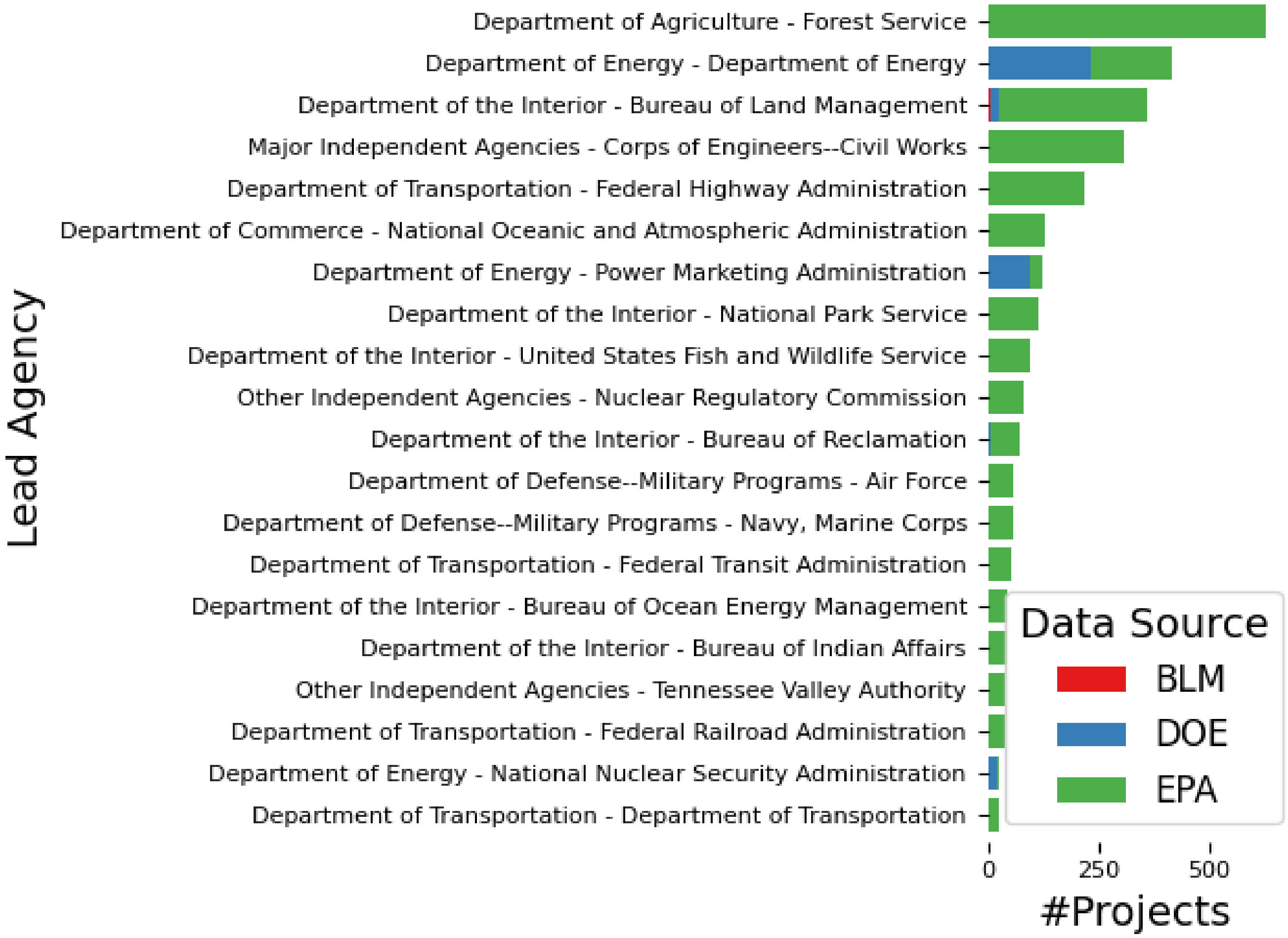


**Figure 11.** Distribution of EIS projects across lead agencies. For brevity, we only visualize the Top-20 most frequent lead agencies in EIS documents from the NEPATECv2 dataset.

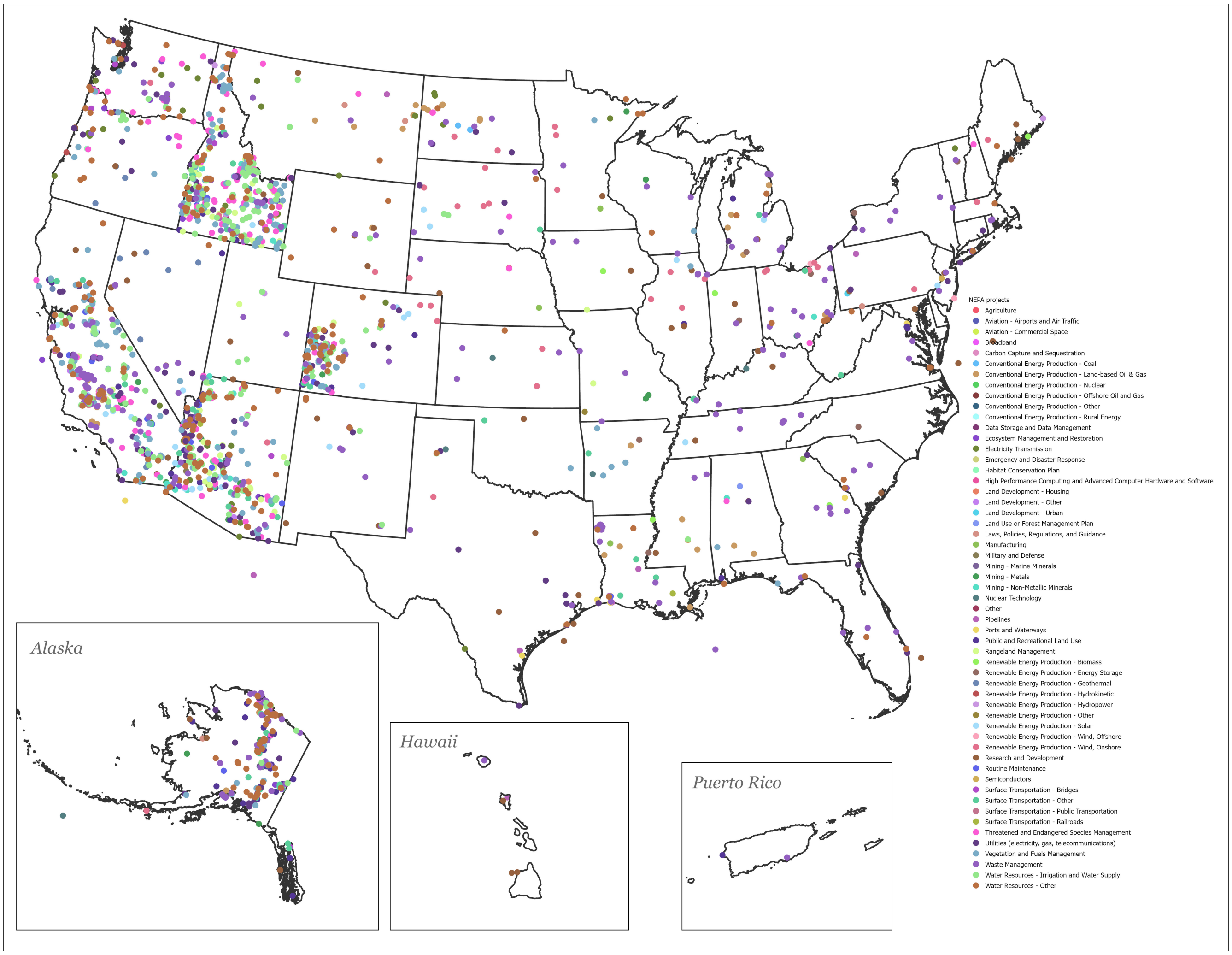

Alaska
Hawaii
Puerto Rico
NEPA projects
● Agriculture
● Aviation - Airports and Air Traffic
● Aviation - Commercial Space
● Broadband
● Carbon Capture and Sequestration
● Conventional Energy Production - Coal
● Conventional Energy Production - Land-based Oil & Gas
● Conventional Energy Production - Nuclear
● Conventional Energy Production - Offshore Oil and Gas
● Conventional Energy Production - Other
● Conventional Energy Production - Rural Energy
● Data Storage and Data Management
● Ecosystem Management and Restoration
● Electricity Transmission
● Emergency and Disaster Response
● Habitat Conservation Plan
● High Performance Computing and Advanced Computer Hardware and Software
● Land Development - Housing
● Land Development - Other
● Land Development - Urban
● Land Use or Forest Management Plan
● Laws, Policies, Regulations, and Guidance
● Manufacturing
● Military and Defense
● Mining - Marine Minerals
● Mining - Metals
● Mining - Non-Metallic Minerals
● Nuclear Technology
● Other
● Pipelines
● Ports and Waterways
● Public and Recreational Land Use
● Rangeland Management
● Renewable Energy Production - Biomass
● Renewable Energy Production - Energy Storage
● Renewable Energy Production - Geothermal
● Renewable Energy Production - Hydrokinetic
● Renewable Energy Production - Hydropower
● Renewable Energy Production - Other
● Renewable Energy Production - Solar
● Renewable Energy Production - Wind, Offshore
● Renewable Energy Production - Wind, Onshore
● Research and Development
● Routine Maintenance
● Semiconductors
● Surface Transportation - Bridges
● Surface Transportation - Other
● Surface Transportation - Public Transportation
● Surface Transportation - Railroads
● Threatened and Endangered Species Management
● Utilities (electricity, gas, telecommunications)
● Vegetation and Fuels Management
● Waste Management
● Water Resources - Irrigation and Water Supply
● Water Resources - Other

**Figure 12.** Geographic distributions of NEPATEC2.0 project locations across the United States from the EA documents. Marker points show exact project centroids by the county level when supported. Marker colors show the project types.

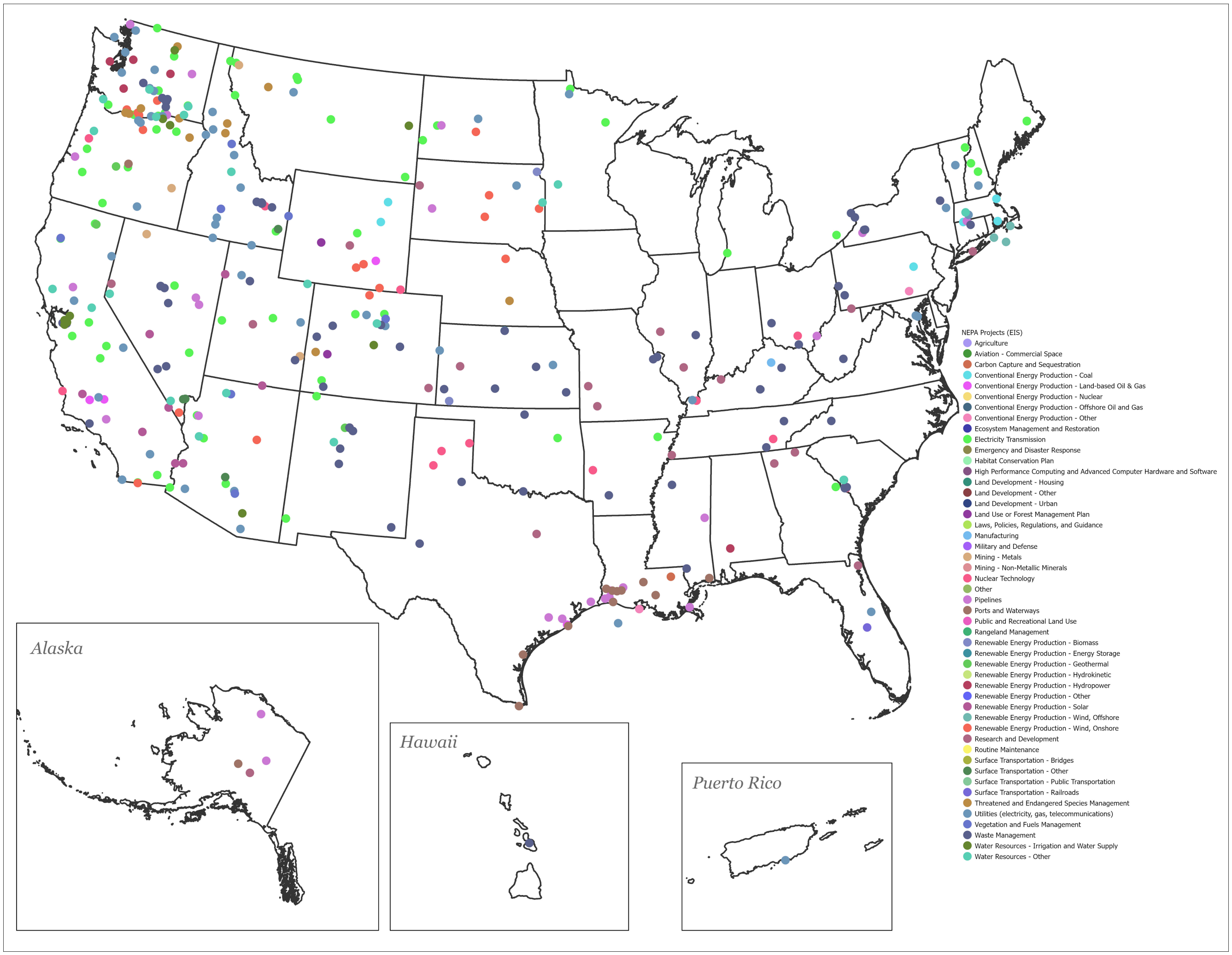

Alaska
Hawaii
Puerto Rico
NEPA Projects (EIS)
● Agriculture
● Aviation - Commercial Space
● Carbon Capture and Sequestration
● Conventional Energy Production - Coal
● Conventional Energy Production - Land-based Oil & Gas
● Conventional Energy Production - Nuclear
● Conventional Energy Production - Offshore Oil and Gas
● Conventional Energy Production - Other
● Ecosystem Management and Restoration
● Electricity Transmission
● Emergency and Disaster Response
● Habitat Conservation Plan
● High Performance Computing and Advanced Computer Hardware and Software
● Land Development - Housing
● Land Development - Other
● Land Development - Urban
● Land Use or Forest Management Plan
● Laws, Policies, Regulations, and Guidance
● Manufacturing
● Military and Defense
● Mining - Metals
● Mining - Non-Metallic Minerals
● Nuclear Technology
● Other
● Pipelines
● Ports and Waterways
● Public and Recreational Land Use
● Rangeland Management
● Renewable Energy Production - Biomass
● Renewable Energy Production - Energy Storage
● Renewable Energy Production - Geothermal
● Renewable Energy Production - Hydrokinetic
● Renewable Energy Production - Hydropower
● Renewable Energy Production - Other
● Renewable Energy Production - Solar
● Renewable Energy Production - Wind, Offshore
● Renewable Energy Production - Wind, Onshore
● Research and Development
● Routine Maintenance
● Surface Transportation - Bridges
● Surface Transportation - Other
● Surface Transportation - Public Transportation
● Surface Transportation - Railroads
● Threatened and Endangered Species Management
● Utilities (electricity, gas, telecommunications)
● Vegetation and Fuels Management
● Waste Management
● Water Resources - Irrigation and Water Supply
● Water Resources - Other

**Figure 13.** Geographic distributions of NEPATEC2.0 project locations across the United States from the EIS documents. Marker points show exact project centroids by the county level when supported. Marker colors show the project types.

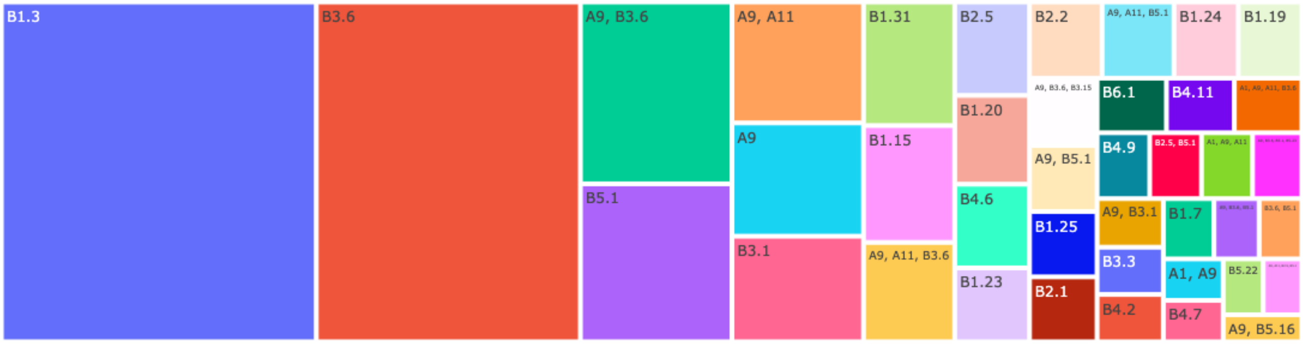


**Figure 14.** Distribution of CE projects across CE categories in DOE CE documents from the NEPATECv2 dataset. IN this treee map, tiles are sized proportionally to the number of projects in each category. For projects with multiple CE categories assigned, groups are counted instead of individual CE categories.

## C Appendix: Metadata Instruction Prompts used in the Metadata Extraction Framework

We used a variety of prompts to extract the metadata from the documents. These are all listed below:

### C.1 Metadata Instruction Prompts for CE Documents

**Metadata Instruction Prompt For Lead Agency**

Your task is to analyze the text and identify one of the Agency Name it belongs to given set of categories.
Below are the list of agencies:
Executive Office of the President Government Sponsored Enterprises International Assistance Programs Financing Vehicles and the Board of Governors of the Federal Reserve General Services Administration Department of Commerce Other Independent Agencies Judicial Branch Department of State Department of Energy Department of Veterans Affairs Department of Labor Legislative Branch Department of Agriculture Department of Education Department of the Interior Department of the Treasury Department of Defense–Military Programs Department of Homeland Security Department of Housing and Urban Development Department of Justice Department of Health and Human Services Department of Transportation Major Independent Agencies

**Metadata Instruction Prompt For Lead Agency (BLM)**

Your task is to analyze the text and identify one of the Agency Name it belongs to given set of categories.
Below are the list of agencies:
Executive Office of the President
Government Sponsored Enterprises
International Assistance Programs
Financing Vehicles and the Board of Governors of the Federal Reserve
General Services Administration
Department of Commerce
Other Independent Agencies
Judicial Branch
Department of State
Department of Energy
Department of Veterans Affairs
Department of Labor
Legislative Branch
Department of Agriculture
Department of Education
Department of the Interior
Department of the Treasury
Department of Defense–Military Programs
Department of Homeland Security
Department of Housing and Urban Development
Department of Justice
Department of Health and Human Services
Department of Transportation
Major Independent Agencies

### Metadata Instruction Prompt For Project Title

You are given a document (PDF) that may include a project description. Your task is to extract the project title. A project title is typically a short phrase that succinctly names the project.
Instructions:
1. Search the document for any direct label indicating a title, such as: - Project Title - Title - Proposed Action - Activity Title - Document Title - Proposed Action Title - Brief Title of Proposed Action
2. If such a label is found, extract the corresponding text as the project title.
3. If no explicit label is found, infer the most appropriate project title from the text by identifying short, descriptive phrases that match the style and structure of titles in the training examples provided (e.g., specific actions, technical projects, construction tasks, research initiatives).
4. The extracted title should be: - Concise (typically 5–15 words) - Specific to the project's objective, location, or scope - Not a full sentence or paragraph
Examples of Valid Titles: - "Replace Obsolete RW Injection Pump Vibration Transmitters" - "Subsea Produced Water Sensor Development" - "Dupee Valley Reserve Property Acquisition and Stewardship Funding" - "Development of Lighting Application Efficacy Measurement Framework" - "Energy Efficiency and Conservation Block Grant Program"
If needed, rephrase or synthesize a logical title based on context and description.

### Metadata Instruction Prompt For Project Title (BLM)

You are given a document (PDF) that may include a project description. Your task is to extract the project title. A project title is typically a short phrase that succinctly names the project.
Instructions:
1. Search the document for any direct label indicating a title, such as: - Project Title - Title - Case Title - Proposed Action - Activity Title - Document Title - Proposed Action Title - Brief Title of Proposed Action
2. If such a label is found, extract the corresponding text as the project title.
3. If no explicit label is found, infer the most appropriate project title from the text by identifying short, descriptive phrases that match the style and structure of titles in the training examples provided (e.g., specific actions, technical projects, construction tasks, research initiatives).
4. The extracted title should be: - Concise (typically 5–15 words) - Specific to the project's objective, location, or scope - Not a full sentence or paragraph
Examples of Valid Titles: - "Replace Obsolete RW Injection Pump Vibration Transmitters" - "Subsea Produced Water Sensor Development" - "Dupee Valley Reserve Property Acquisition and Stewardship Funding" - "Development of Lighting Application Efficacy Measurement Framework" - "Energy Efficiency and Conservation Block Grant Program"
If needed, rephrase or synthesize a logical title based on context and description.

### Metadata Instruction Prompt For Project Location

You are given a document (PDF) that may include a project description. Your task is to extract the most likely project location in the form of County, City, and State. Follow the guidelines below:
1. Primary Extraction: Search the text for explicit mentions of project location. Look for County, City, and State names directly associated with the project activity.
2. Multiple Mentions: If multiple locations are mentioned, return the one that appears to be most directly related to the project based on proximity to project-specific terms (e.g., "project located in...", "construction in...", etc.).
3. Fallback to Institution: If no County, City, or State is explicitly mentioned, but an institution or organization (e.g., a university, DOE lab, or government agency) is named, use your knowledge base to infer the most likely location of that institution.
4. Unknown Location: If you cannot determine any County, City, or State from the text or institution, return "UNK".
Format your output as: <County>, <City>, <State> Leave out components not present in the text. Preserve the original form and casing as mentioned in the document.

## Metadata Instruction Prompt For Project Location (BLM)

Given the details of a project, your task is to extract the **project location**. Follow the hierarchical extraction logic below:
1. **Primary Extraction – Legal Land Description (Most Precise):** - First, check for structured land description formats typically used in federal or state land documents. These may include keywords like: - **Meridian**, **Township**, **Range**, **Section**, **Lot**, **Quarter**, or **Legal Description** - These are often listed together in a block and may appear under headings like **Location**, **Project Location**, **Location of the Proposed Action**, or within the main project description. - Extract the **entire contiguous block** that forms the land description (e.g., *"[Meridian], [State], Township X [North/South], Range Y [East/West], Section Z, Lot N"*).

- Preserve formatting, line breaks, punctuation, and casing exactly as in the source, except for example like where the locations is not points to any *STATE*. In such cases try to find out state, county form documents and append that to the extracted location.
- For example if the location is something like "T.09S., R.02W., Section 26 (see attached map)" which doesnot mention the **STATE** but point out the location of the map, it such case look for the location in map or infer city, county, state from the project descriptions.

2. **Secondary Extraction – Labeled Project Location:** - If no legal land description is found, look for labeled fields such as: - **Location:**, **Project Location:**, **Site Location:**, **Location of the Proposed Action:**, or similar - Extract the full value next to or beneath that label, keeping the original formatting.
3. **Contextual Mentions – Proximity to Project Terms:** - If multiple locations are mentioned, prefer the one: - Near the project name or case number - Mentioned with phrases like *“project is located at…”*, *“site of action…”*, *“land described as…”*
4. **Fallback to Institution:** - If no project-specific location is found, but a named institution or agency is mentioned (e.g., a federal lab, state office, university), use known institutional headquarters to infer its likely location.
5. **Unknown Location:** - If you cannot confidently determine a location using the above rules, return ‘"UNK"‘.
6. **Multiple Mentions:** - If multiple locations are mentioned, return the one that appears to be most directly related to the project based on proximity to project-specific terms (e.g., "project located in...", "construction in...", etc.).

Note: - Do not return multiple locations. - Do **not** extract broader administrative areas (like county or state alone) unless no more specific location is available. - Do **not** infer or summarize — extract only what is explicitly or structurally present in the document.

## Metadata Instruction Prompt For Project Sponsor

You are given a document (PDF) that may include a project description. Your task is to extract the *project sponsor*—typically the organization, agency, university, or institution responsible for funding, managing, or overseeing the project.

Your goal is to extract the sponsor name with high precision. Follow these detailed instructions:

Instructions:

1. **Prioritize Labeled Fields (High Confidence)** Look for labeled fields or sections that strongly suggest the sponsor, including but not limited to: - "Recipient" or "Recipient Name", if "Sub-recipients" is also present then report Recipient field information (high priority ) and ignore the other field. - "Program or Field Office" - "Program or Program Office" - "Organization Name" - "Project Sponsor" - "Header" or "Page Header"
2. **Secondary Sections (Supporting Information)** If labeled fields are not definitive, review these contextual sections for sponsor indicators: - "Description of the Proposed Action" - "Proposed Action Title" - "Brief Description of Proposal" - "Project Description" or "Project Description and Purpose" - "Section B. Project Description" - "Project Contact" - "Proposed Action Description"
3. **Inferred Sponsors (If Labels Missing)** - Identify entities that commonly act as sponsors using context (e.g., DOE labs, universities, private companies, government offices). - If multiple entities are mentioned, prefer the one aligned with funding or administrative oversight.
4. **Disambiguation and Normalization Rules** Apply these rules to improve accuracy: - If "NETL", "U.S. Department of Energy - NETL", or "National Energy Technology Laboratory (NETL)" appears, normalize to **"National Energy Technology Laboratory (NETL)"**. - If the entity is "Savannah River Site" but context implies DOE management, accept **"Savannah River Site"** unless DOE is more explicit. - Normalize short names like "Texas A&M" to full institutional names when possible, e.g., **"Texas A&M University"**. - For national labs or DOE field offices, prefer the specific entity (e.g., "Idaho National Laboratory") over the general "Department of Energy" unless only DOE is stated.
5. **Format Guidelines** - Return only the name of the organization, lab, university, agency, or company. - Do **not** return personal names, job titles, or long paragraphs. - If multiple candidates are present, pick the one with the strongest contextual or labeled association to the project scope.
6. **Avoid** - General phrases like "Department of Energy" if a more specific sub-entity (e.g., "Savannah River Site") is mentioned nearby.

### Metadata Instruction Prompt For Project Sponsor (BLM)

You are given a document (PDF) that may include a project description. Your task is to extract the *project sponsor*—typically the organization, agency, university, or institution responsible for funding, managing, or overseeing the project.
Your goal is to extract the sponsor name with high precision. Follow these detailed instructions:
Instructions: 1. **Common Sponsor Locations in the order of preference** (High-Confidence) - "Proponent" - "Applicant" - "Description of Proposed Action" - "Project Description" - "Project Contact" - In these sections look for the agency or company name (preferred) as the sponsor.
2. **Contextual Inference** (If sponsor not found in common sponsor locations) (Low-confidence) - Use contextual cues to infer the most likely sponsor. - Identify organizations mentioned in administrative, funding, or managerial roles. - If multiple entities appear, choose the one most aligned with project authority or funding.
3. **Other Rules** - In the sponsor locations if company name and agency name both are mentioned, typically it will be the company that is the sponsor. - Prefer the **specific lab or field office** over a broad parent agency unless specificity is not available. - Output only the sponsor organization name—do not include individuals, roles, or extended context. - If multiple candidates are present, select the one with the strongest contextual link to oversight or funding. - Avoid “Bureau of Land Management”, "Department of Agency" like agencies if sub-entity is mentioned. - if you are absolutely certain that sponsor information like, name of a person, owner, company, agency nothing like this information present then text then return ’UNK’

### Metadata Instruction Prompt For Project Type

Based on the description of the proposed action outlined in the project documents, carefully review the provided list of project types.

Select one or more project types that most accurately reflect the nature and scope of the proposed action.

**CRITICAL CONSTRAINT**: You MUST choose ONLY from this exact list - do not modify, abbreviate, or create variations of these categories:

[’Aviation - Airports and Air Traffic’, ’Surface Transportation - Public Transportation’, ’Manufacturing’, ’Rangeland Management’, ’Surface Transportation - Bridges’, ’Renewable Energy Production - Energy Storage’, ’Research and Development’, ’Ecosystem Management and Restoration’, ’Waste Management’, ’Laws, Policies, Regulations, and Guidance’, ’Renewable Energy Production - Solar’, ’Mining - Marine Minerals’, ’Vegetation and Fuels Management’, ’Renewable Energy Production - Biomass’, ’Land Development - Other’, ’Mining - Metals’, ’Water Resources - Other’, ’Cybersecurity’, ’Conventional Energy Production - Rural Energy’, ’Conventional Energy Production - Land-based Oil & Gas’, ’Land Development - Urban’, ’Conventional Energy Production - Offshore Oil and Gas’, ’High Performance Computing and Advanced Computer Hardware and Software’, ’Emergency and Disaster Response’, ’Public and Recreational Land Use’, ’Aviation - Commercial Space’, ’Conventional Energy Production - Nuclear’, ’Renewable Energy Production - Geothermal’, ’Surface Transportation - Other’, ’Renewable Energy Production - Hydropower’, ’Military and Defense’, ’Renewable Energy Production - Other’, ’Renewable Energy Production - Hydrokinetic’, ’Land Use or Forest Management Plan’, ’Semiconductors’, ’Ports and Waterways’, ’Surface Transportation - Railroads’, ’Artificial Intelligence and Machine Learning’, ’Broadband’, ’Conventional Energy Production - Other’, ’Mining - Non-Metallic Minerals’, ’Renewable Energy Production - Wind, Onshore’, ’Pipelines’, ’Other’, ’Quantum Information Science and Technology’, ’Routine Maintenance’, ’Utilities (electricity, gas, telecommunications)’, ’Nuclear Technology’, ’Land Development - Housing’, ’Renewable Energy Production - Wind, Offshore’, ’Data Storage and Data Management’, ’Carbon Capture and Sequestration’, ’Electricity Transmission’, ’Water Resources - Irrigation and Water Supply’, ’Habitat Conservation Plan’, ’Agriculture’, ’Threatened and Endangered Species Management’, ’Conventional Energy Production - Coal’]

Scope of Review: - Use **only** the information from the following sections: - *Project title* - *Document title* - *Proposed action* - *Project description* - **Ignore** text related to categorical exclusions, determination justifications, regulatory references, and any content that appears **after** the project description, including any mention of routine maintenance or NEPA compliance language.

Classification Guidelines: - Assign multiple project types if the project involves multiple qualifying areas, but never invent new ones. *Assign at least three relevant project types if applicable.* - If the action involves **water infrastructure, systems, or equipment** (e.g., water return pumps, cooling systems, water distribution or treatment systems), classify as **Water Resources** even if the work is a repair, upgrade, or replacement. - If the project involves **tritium**, **radiological materials**, or **nuclear activities**, assign a **Nuclear** category along with other relevant types. - If the proposed action involves **handling, converting, recovering, removing, or treating waste** (solid, hazardous, radioactive, thermal, electric, gaseous, or wastewater), **always** include **Waste Management** as a project type. This applies even if the waste is being reused or converted (e.g., waste heat recovery). - Projects that utilize or repurpose **vehicle or industrial waste heat** for conversion into electricity or other uses should be classified under **Waste Management**, alongside any energy-related categories that may also apply. - Renewable energy also is a form of waste management, so if the project involves **renewable energy production** (solar, wind, hydro, biomass, geothermal, hydrokinetic, energy storage), assign the appropriate **Renewable Energy Production** category along with **Waste Management**. - If the action involves **leasing, developing, or using land** for residential, commercial, industrial, or infrastructure purposes, assign appropriate **Land Development** categories along with other relevant types. - If the project involves **electric power systems** (generation, storage, distribution, or transmission), include "Electricity Transmission" or similar along with other relevant types. - If the project mentions **water-related activities**, such as cooling water systems, water resources infrastructure, or hydrologic management, assign **Water Resources** even if the action is repair, maintenance, or equipment replacement. Put this along other relevant types. - If the project involves manufacturing then assign **Manufacturing** among other relevant types. - If the project involves rare earth elements then mining category is applicable among other relevant types.

Important Considerations: - Do **not** default to "Routine Maintenance" just because it is listed as the Categorical Exclusion type. Instead, classify based on the **actual nature of the project work**, especially from the project title and description. - Prefer **specific project types** (e.g., "Vegetation and Fuels Management", "Renewable Energy Production - Solar", "Electricity Transmission", "Manufacturing") over general categories (e.g., "Routine Maintenance", "Research and Development"). - If multiple project types apply, prefer the **most specific** or **environmentally significant** category over general ones like "Routine Maintenance" or "Research and Development". - Use "Routine Maintenance" or "Research and Development" only as a **last resort**, when the project has **no suitable classifications**. If, and only if, none of the listed project types are appropriate, choose ’Other’.

Your classification should reflect an accurate environmental lens on the major proposed activity or infrastructure.

### Metadata Instruction Prompt For Project Type (BLM)

Based on the description of the proposed action outlined in the project documents, carefully review the provided list of project types. Select one or more project types that most accurately reflect the nature and scope of the proposed action.

**CRITICAL CONSTRAINT**: You MUST choose ONLY from this exact list - do not modify, abbreviate, or create variations of these categories:

[’Aviation - Airports and Air Traffic’, ’Surface Transportation - Public Transportation’, ’Manufacturing’, ’Rangeland Management’, ’Surface Transportation - Bridges’, ’Renewable Energy Production - Energy Storage’, ’Research and Development’, ’Ecosystem Management and Restoration’, ’Waste Management’, ’Laws, Policies, Regulations, and Guidance’, ’Renewable Energy Production - Solar’, ’Mining - Marine Minerals’, ’Vegetation and Fuels Management’, ’Renewable Energy Production - Biomass’, ’Land Development - Other’, ’Mining - Metals’, ’Water Resources - Other’, ’Cybersecurity’, ’Conventional Energy Production - Rural Energy’, ’Conventional Energy Production - Land-based Oil & Gas’, ’Land Development - Urban’, ’Conventional Energy Production - Offshore Oil and Gas’, ’High Performance Computing and Advanced Computer Hardware and Software’, ’Emergency and Disaster Response’, ’Public and Recreational Land Use’, ’Aviation - Commercial Space’, ’Conventional Energy Production - Nuclear’, ’Renewable Energy Production - Geothermal’, ’Surface Transportation - Other’, ’Renewable Energy Production - Hydropower’, ’Military and Defense’, ’Renewable Energy Production - Other’, ’Renewable Energy Production - Hydrokinetic’, ’Land Use or Forest Management Plan’, ’Semiconductors’, ’Ports and Waterways’, ’Surface Transportation - Railroads’, ’Artificial Intelligence and Machine Learning’, ’Broadband’, ’Conventional Energy Production - Other’, ’Mining - Non-Metallic Minerals’, ’Renewable Energy Production - Wind, Onshore’, ’Pipelines’, ’Other’, ’Quantum Information Science and Technology’, ’Routine Maintenance’, ’Utilities (electricity, gas, telecommunications)’, ’Nuclear Technology’, ’Land Development - Housing’, ’Renewable Energy Production - Wind, Offshore’, ’Data Storage and Data Management’, ’Carbon Capture and Sequestration’, ’Electricity Transmission’, ’Water Resources - Irrigation and Water Supply’, ’Habitat Conservation Plan’, ’Agriculture’, ’Threatened and Endangered Species Management’, ’Conventional Energy Production - Coal’]

Scope of Review: - Use **only** the information from the following sections: - *Project title* - *Document title* - *Proposed action* - *Project description* - **Ignore** text related to categorical exclusions, determination justifications, regulatory references, and any content that appears **after** the project description, including any mention of routine maintenance or NEPA compliance language.

Classification Guidelines: - Assign multiple project types if the project involves multiple qualifying areas, but never invent new ones. *Assign at least three relevant project types if applicable.* - If the action involves **water infrastructure, systems, or equipment** (e.g., water return pumps, cooling systems, water distribution or treatment systems), classify as **Water Resources** even if the work is a repair, upgrade, or replacement. - If the project involves **tritium**, **radiological materials**, or **nuclear activities**, assign a **Nuclear** category along with other relevant types. - If the proposed action involves **handling, converting, recovering, removing, or treating waste** (solid, hazardous, radioactive, thermal, electric, gaseous, or wastewater), **always** include **Waste Management** as a project type. This applies even if the waste is being reused or converted (e.g., waste heat recovery). - Projects that utilize or repurpose **vehicle or industrial waste heat** for conversion into electricity or other uses should be classified under **Waste Management**, alongside any energy-related categories that may also apply. - Renewable energy also is a form of waste management, so if the project involves **renewable energy production** (solar, wind, hydro, biomass, geothermal, hydrokinetic, energy storage), assign the appropriate **Renewable Energy Production** category along with **Waste Management**. - If the action involves **leasing, developing, or using land** for residential, commercial, industrial, or infrastructure purposes, assign appropriate **Land Development** categories along with other relevant types. - If the project involves **electric power systems** (generation, storage, distribution, or transmission), include "Electricity Transmission" or similar along with other relevant types. - If the project mentions **water-related activities**, such as cooling water systems, water resources infrastructure, or hydrologic management, assign **Water Resources** even if the action is repair, maintenance, or equipment replacement. Put this along other relevant types. - If the project involves manufacturing then assign **Manufacturing** among other relevant types. - If the project involves rare earth elements then mining category is applicable among other relevant types.

Important Considerations: - Do **not** default to "Routine Maintenance" just because it is listed as the Categorical Exclusion type. Instead, classify based on the **actual nature of the project work**, especially from the project title and description. - Prefer **specific project types** (e.g., "Vegetation and Fuels Management", "Renewable Energy Production - Solar", "Electricity Transmission", "Manufacturing") over general categories (e.g., "Routine Maintenance", "Research and Development"). - If multiple project types apply, prefer the **most specific** or **environmentally significant** category over general ones like "Routine Maintenance" or "Research and Development". - Use "Routine Maintenance" or "Research and Development" only as a **last resort**, when the project has **no suitable classifications**. If, and only if, none of the listed project types are appropriate, choose ’Other’.

Your classification should reflect an accurate environmental lens on the major proposed activity or infrastructure.

### Metadata Instruction Prompt For Document Title

You are given a document (PDF) that may include a formal report or project description. Text is obtained from a combination of OCR, fillable form text, and default PDF reader extraction. Your task is to extract the **document title** which typically appears at the top or in the header of the document—within the **first few lines** of the extracted text.
Instructions:
1. In most cases, the document title appears as fist or second line of text. Do not use project title if you find relevant text in the first line of OCR extracted text. (High confidence) 2. If no clear header is found, use a recognizable document identifier (e.g., “Categorical Exclusion Determination Form”, “NEPA Determination”, or similar boilerplate document phrases) (Low confidence).
The extracted title should: - Be concise (2-15 words)
Examples of valid document titles: - "U.S. Department of Energy Categorical Exclusion Determination Form" - "Development of Thermal Breakout Technology for Determining In Situ Stress" - "Environmental Review for Categorical Exclusion Determination" - "DWPF Training Campus Complex Expansion" - "Energy Efficiency and Conservation Block Grant Program"

### Metadata Instruction Prompt For Document Title (BLM)

You are given documents that may include a formal report or project description. Your task is to extract the **document title** which typically appears at the top or in the header of the document—within the **about first four lines** of the extracted text.
Instructions:
1. In most cases, the document title appears as fist or second line of text. Do not use project title if you find relevant text in the first line extracted text. 2. If no clear header is found, use a recognizable document identifier (e.g., “Categorical Exclusion Determination Form”, “NEPA Determination”, Categorical Exclusion followed by project number/title, or similar boilerplate document phrases).
The extracted title should: - Be concise (single line).
Examples of valid document titles: - "Categorical Exclusion Determination Form" - "Categorical exlucsion [project number] [title]" - "Environmental Review for Categorical Exclusion Determination"

### Metadata Instruction Prompt For Prepared By

Text is obtained from a combination of OCR, fillable form text, and default PDF reader extraction.
Identify the agency or organization(s) responsible for preparing the document. These are typically found in the document header, metadata, form fields, or a “list of preparers” section, and usually include a federal agency, contractor, or project proponent.
Only extract institutional affiliations (e.g., Department of Energy, Environmental Protection Agency, Forest Service) — do not include individual names.
If the same agency is repeated multiple times (e.g., “Department of Energy - Department of Energy”), treat it as a single entry.
Sub-agencies (e.g., "Energy Programs", "National Nuclear Security Administration") may be included if they add meaningful specificity.
Do not list duplicates or placeholder values such as "0".
Return a concise list of unique responsible organizations.

### Metadata Instruction Prompt For Prepared By (BLM)

Text is obtained from a combination of OCR, fillable form text, and default PDF reader extraction.
Identify the agency or organization(s) responsible for preparing the document. These are typically found in the document header, metadata, form fields, or a “list of preparers” section, and usually include a federal agency, contractor, or project proponent.
Only extract institutional affiliations (e.g., Department of Energy, Environmental Protection Agency, Forest Service) — do not include individual names.
If the same agency is repeated multiple times (e.g., “Department of Energy - Department of Energy”), treat it as a single entry.
Sub-agencies (e.g., "Energy Programs", "National Nuclear Security Administration") may be included if they add meaningful specificity.
Do not list duplicates or placeholder values such as "0".
Return a concise list of unique responsible organizations.

## Metadata Instruction Prompt For CE category

Read the PDF document and identify all categorical exclusion codes.

Only focus on codes that: 1. Are preceded by a checkbox that is **explicitly marked as selected** (e.g., with a checkmark, cross, or filled box). 2. Sometimes just mentioned under categorical exlusion as list without if there are are not selection options. 3. Begin with the letter **A** or **B**

Ignore all other content in the document.

For each selected code, extract: - The **code** (e.g., B1.3 of the text B1.3 - Routine Maintenance) only. Don’t need to include text description of the code. - For multiple codes, put them in commma separated list.

Return only the list of selected A/B categorical exclusion codes. Do **not** include unselected codes or unrelated text from the PDF.

## Metadata Instruction Prompt For CE category (BLM)

Below are the extracted text of the the pdf documents <Filename> Filename of the PDF document. </Filename> <Text> Extracted text from the PDF document. This typically includes the first 1–2 pages, obtained via PDF text extraction tools. </Text>

From the extracted text identify all Categorical Exclusion that were applied for the projects. Here are the Guidelines for them:

- They text usually mentions something like "The Proposed Action is categorically excluded from further documentation under the National Environmental Policy Act (NEPA)" in accordance with .... some details.

An example text wehre Categorical exclusion information would be, "The Proposed Action is categorically excluded from further documentation under the National Environmental Policy Act (NEPA) in accordance with 516 DM 2, Appendix 1, or 516 DM 11.9. Appendix 4, Departmental Categorical Exclusions:

F. Solid Minerals, 10. Disposal of mineral materials, such as sand, stone, gravel, pumice, pumicite, cinders, and clay, in amounts not exceeding 50,000 cubic yards or disturbing more than 5 acres, except in riparian areas."

For the above case, output this in following way:

<source> 516 DM 2, Appendix 1, or 516 DM 11.9. Appendix 4, Departmental Categorical Exclusions </source> <code> F. Solid Minerals, 10. Disposal of mineral materials, such as sand, stone, gravel, pumice, pumicite, cinders, and clay, in amounts not exceeding 50,000 cubic yards or disturbing more than 5 acres, except in riparian areas. </code>

- If source is not present, the put <source>None</source><code>text for the code</code>
- Sometimes the CE codes and source are mentioned in the rational sections.
- Sometimes it is mention in the part where it says, the applicable Categorical Exclusion reference are ...
- For multiple categorical exclusions, put them in the above formated list.

## Metadata Instruction Prompt For CE category (USDA)

Below are the extracted texts from PDF documents, Your task is to identify all **Categorical Exclusion (CE)** references that were applied to the proposed project described in the extracted text. These are usually legal codes or regulatory citations that justify excluding the project from further environmental documentation under NEPA.

### Guidelines:

- Look for any statements indicating that the project is *categorically excluded* from an Environmental Assessment (EA) or Environmental Impact Statement (EIS) under NEPA. These often appear in phrases like: - "This action is categorically excluded..." - "The proposal qualifies for a categorical exclusion..." - "In accordance with [source], this action falls under CE..." - "Applicable category: [source]..."
- Each CE reference typically has two components: 1. **Source**: A regulation or code identifying the authority or section (e.g., ‘36 CFR 220.6(e)(15)‘, ‘516 DM 11.9‘, etc.) 2. **Code Text/Description**: A specific clause or description of the CE (e.g., "Issuance of a new special use authorization...", "Replacing an underground cable trunk...").
- **Format your output** like this for each CE found:

<source> [Regulatory source or reference] </source> <code> [Full text or description of the CE category that applies to the project] </code>

## Metadata Instruction Prompt For Project Description

Read the PDF document and extract the project description. These are typically paragraphs of text describing the details of projects. Look for sections or labels that says description. Start the answer by naming that label and then the text. Only use the text from the pdf do not add anything. In special cases, the description might need to consider the text falls in rational label as a supporting additional description. In such case add both of those text with their labels as answer.

### Metadata Instruction Prompt For Project Description (BLM)

Below are the extracted text of the the pdf documents <Filename> Filename of the PDF document. </Filename> <Text> Extracted text from the PDF document. </Text>
From the extracted text read extract the project description. These are typically paragraphs of text describing the details of projects. Look for sections or labels that says description or proposed action, for example: **Description of Proposed Action**, **Description of Project**, **Project Description**, **Proposed Action**, or similar to these.
Start the answer by naming that label and then the text. Only use the text from the pdf do not add anything. In special cases, the description might need to consider the text falls in rational label as a supporting additional description. In such case add all of those sections and their corresponding descriptions as answer. Formatting as, **Section name: Detail project description** as is in the text. Do not modify anything.

### Metadata Instruction Prompt For Section Title

Extract the **document section or form title** from this file.
- The title typically appears at the top or in the header of the document—within the **first few lines** of the extracted text. - Text is obtained from a combination of OCR, fillable form text, and default PDF reader extraction. - The goal is to identify **formal section titles or document headers**, such as: - “Categorical Exclusion Determination Form” - “U.S. Department of Energy Categorical Exclusion Determination” - “Environmental Review for Categorical Exclusion Determination”
The extracted title should: - Be concise (5–15 words) - Clearly describe the document’s purpose or subject - Preserve known or standard titles (e.g., NEPA forms), even if they include generic terms - Avoid full sentences unless part of an official form title
Return only the most relevant document/form/section title.

### Metadata Instruction Prompt For Section Title (BLM)

Extract the **document section or form title** from this file.
- The title typically appears at the top or in the header of the document—within the **first few lines** of the extracted text. - The goal is to identify **formal section titles or document headers**, such as: - “Categorical Exclusion Determination Form” - “Environmental Review for Categorical Exclusion Determination” - [Title] Categorical Exlclusion [Some other details] - "Decision for Categorical Exclusion [Project number and others]"
The extracted title should: - Be concise (5–15 words) - Clearly describe the document’s purpose or subject - Preserve known or standard titles (e.g., NEPA forms), even if they include generic terms - Avoid full sentences unless part of an official form title
Return only the most relevant document/form/section title.

## C.2 Metadata Instruction Prompts for EA Documents

### Metadata Instruction Prompt For Main Document

You are tasked with predicting whether a file contains all or part of a main document, based on the file name and it’s contents. The file is not a part of a main document if it contains only appendices, letter, addendum or other supporting information. If a file is prepared in support of another file, then it is also not considered as part of the main document, except when the file itself represents a distinct document type (e.g., Supplemental Analysis) or, in the case of addenda, when it contains a table of contents and a substantial number of pages, in which case it should be considered a main document on its own.
The main document is defined as the document title page and summary / executive summary through all chapters, but excludes appendices. Reviewing the file’s table of contents and/or first and last section headings may be helpful in making this determination.

### Metadata Instruction Prompt For Volume Or Section Title

You are tasked with extracting the complete title of a document section or volume from a text file that is part of an Environmental Assessment. The title can consist of multiple components, such as the project name, section/volume designations (e.g., Appendix A, Chapter 5), and descriptive names (e.g., Karst Characterization Studies). A title must contain at-least one of these components.
# Instructions: 1. Identify the complete title, ensuring it includes all relevant components (i.e. project name, section/volume designation, descriptive name of volume/section) as they appear in the document. Do not miss any of the components. The components may be separated by (multiple) new lines in the file. 2. Remove any document identification numbers (e.g., codes like DOE/EA-XXXX, EIS-XXXX, or similar alphanumeric references) that may appear as part of the heading. These are not considered part of the section or volume title. 3. Concatenate the components of the extracted title into a single line, ensuring they are separated by only a single space. There should be no new line characters or multiple consecutive spaces in the final output. 4. Exclude appended location information (e.g., district, forest, county, state) from the end of a heading unless it appears in the middle of the title. 5. Provide a reasoning for the extracted title. If any of the above-mentioned components are missing in the extracted title, provide an explanation for that. 6. Do not truncate or abbreviate any part of the title. 7. Exclude any dates present in the title, unless they appear in the middle of the title. 8. Ensure the extracted title is an exact match to the text within the file, , excluding any omitted document numbers as per instruction #2. 9. If multiple titles are present, return the title that appears earliest in the file. 10. The filename is provided as context but may not directly represent the title.

### Metadata Instruction Prompt For Project Type

You are a vital component of an environmental regulatory compliance system. Your role is to classify projects described in sensitive environmental assessment documents (e.g., EAs) with the utmost precision. Misclassification can trigger serious regulatory, ecological, and public-health consequences.
nEach project description (the ‘InputText‘) outlines a proposed action—its purpose, scope, and technical elements. Your classifications guide oversight, resource allocation, and risk mitigation. Transparency and accuracy in your reasoning are therefore essential.
n
nInstructions:
n1. Read the ‘InputText‘ carefully, identifying the project’s nature, scope, and primary objectives.
n2. In the ‘reasoning‘ section, document a clear, step-by-step justification. Cite specific keywords or passages that led you to each conclusion.
n3. From the provided list of ‘ProjectType‘ options, select *all* categories that accurately reflect the project’s activities.
n a. Where an option is structured as “Primary Category – Subcategory,” treat the first element as the primary category and the second as a subcategory.
n b. If the chosen subcategory appears particularly relevant, also consider and rank any sibling subcategories under the same primary category, if you think they are applicable.
n4. For projects spanning multiple domains (for example, waste management elements within a fossil-fuel facility), include each pertinent category (e.g., “Waste Management” and “Conventional Energy Production – Land-based Oil & Gas”).
n5. If no listed category fits, select ‘Other‘.
n6. Deliver your answer in two parts:
n • reasoning: A numbered, narrative explanation of your decision process.
n • ProjectType: A list of the project types selected.
n
nThe list of project types:

### Metadata Instruction Prompt For Project Title

Extract the title of this project from the provided document. A project title is a descriptive name that represents the project and is often found on the cover page or title page, though it may not always be explicitly labeled. Avoid including document types (e.g., categorical exclusions, environmental assessments). If a clear project title is not identifiable, infer an appropriate title based on the proposed action described in the document.

### Metadata Instruction Prompt For File Type

You are tasked with assigning a type to a file on the contents of the file. The type of the file will be one from the following list: [CE, DEA, EA, DEIS, FEIS, FONSI, ROD, OTHER]. The document type is typically evident from the document cover or title page, summary, or headers/footers. When determining whether the file should be classified as a DEA (Draft Environmental Assessment), you must carefully check the document for the presence of the exact phrase "Draft Environmental Assessment" or other closely similar wordings that indicate this status.
The full meanings of the document types are as follows: CE: Categorical Exclusion. DEA: Draft Environmental Assessment. EA: Environmental Assessment. DEIS: Draft Environmental Impact Statement. FEIS: Final Environmental Impact Statement. FONSI: Finding of No Significant Impact. ROD: Record of Decision. OTHER: Documents that are not of any of the above-mentioned types.

## C.3 Metadata Instruction Prompts for EIS Documents

### Metadata Instruction Prompt For File Type

"You are tasked with assigning a type to a file on the contents of the file. The type of the file will be one from the following list: [CE, DEA, EA, DEIS, FEIS, FONSI, ROD, OTHER]. The document type is typically evident from the document cover or title page, summary, or headers/footers.
The full meanings of the document types are as follows: CE: Categorical Exclusion. DEA: Draft Environmental Assessment. EA: Environmental Assessment. DEIS: Draft Environmental Impact Statement. FEIS: Final Environmental Impact Statement. FONSI: Finding of No Significant Impact. ROD: Record of Decision. OTHER: Documents that are not of any of the above-mentioned types."

### Metadata Instruction Prompt For Location

You are tasked with extracting and providing accurate location information for the study area described in the provided project documents, ensuring consistent formatting across all responses. Your output should also include a single estimated latitude/longitude for the centroid of the study area. If the study area is limited to a single city or small region, include the city or region name, county name, state name, and format the response as: City Name, County Name, State Name (Lat/Lon: [latitude, longitude]), for example: Seattle, King County, WA (Lat/Lon: 47.6062, -122.3321). If the study area spans multiple counties but includes three or fewer, list the county names, state names, and format your response as: County Name, State Name (Lat/Lon: [latitude, longitude]), for example: King County, Garfield County, WA; Lake County, OR(Lat/Lon: 47.57511, -122.32508). For study areas exceeding three counties, do not provide county names, and provide only the state names, along with latitude/longitude and confidence score, formatted as: State Name (Lat/Lon: [latitude, longitude]), such as Washington, Oregon, Idaho (Lat/Lon: 45.5000, -120.5000). For marine or offshore study areas, include a descriptive location along with latitude/longitude in the format: Marine zone located [descriptive location] (Lat/Lon: [latitude, longitude]), for example: Marine zone offshore west of Los Angeles, CA (Lat/Lon: 33.7500, -119.2500). Focus exclusively on the study area described in the documents and ensure unrelated locations mentioned in the text are excluded. All responses must be concise, accurate, and adhere strictly to these formatting rules.

### Metadata Instruction Prompt For Project Title

You are an environmental document specialist. Your task is to identify and extract the precise, formal project title of the provided Environmental Impact Statement. A project title is a descriptive name for a project. Project titles may be written on the cover page or title page of documents available for this project, but are not always explicitly labeled.
Instructions:
1. Ensure the extracted project title does not include any file names, volume numbers, or internal section titles. 2. A project title should be revised to omit the type of document (e.g., categorical exclusions, environmental assessment, environmental impact statement), if present. 3. Do not have abbreviations in the project title, use their full version. 4. If you find multiple acceptable titles at different places, reurn the one that apears earlier in the project text. 5. If a project title cannot be clearly identified from the avialable documents, you may infer a title, typically based on the proposed action described in one or more of the documents (preferentially the most recent).

### Metadata Instruction Prompt For Document Title

You are an environmental document specialist. Your task is to identify and extract the precise, formal title of the provided Environmental Impact Statement. The title is typically found on the cover or initial pages of the document.
Instructions: 1. Ensure the extracted title is the full document title and does not include any file names, volume numbers, or internal section titles. 2. Do not have abbreviations in the document title, use their full version. 3. If you find multiple acceptable titles at different places in the document, return the one that appears earlier in the document.

### Metadata Instruction Prompt For Main Document

You are tasked with predicting whether a file contains all or part of a main document, based on the file name and it's contents. The file is not a part of a main document if it contains only appendices, letter, addendum or other supporting information. If a file is prepared in support of another file, then it is also not considered as part of the main document.
The main document is defined as the document title page and summary / executive summary through all chapters, but excludes appendices. Reviewing the file's table of contents and/or first and last section headings may be helpful in making this determination.

### Metadata Instruction Prompt For Volume Or Section Title

You are tasked with extracting the complete title of a document section or volume from a text file that is part of an Environmental Impact Statement. The title can consist of multiple components, such as the project name, section/volume designations (e.g., Appendix A, Chapter 5), and descriptive names (e.g., Karst Characterization Studies). A title must contain at-least one of these components.
# Instructions: 1. Identify the complete title, ensuring it includes all relevant components (i.e. project name, section/volume designation, descriptive name of volume/section) as they appear in the document. Do not miss any of the components. The components may be separated by (multiple) new lines in the file. 2. Provide a reasoning for the extracted title. If any of the above-mentioned components are missing in the extracted title, provide an explanation for that. 3. Do not truncate or abbreviate any part of the title. 4. Exclude any dates present in the title, unless they appear in the middle of the title. 5. Ensure the extracted title is an exact match to the text within the file. 6. If multiple titles are present, return the title that appears earliest in the file. 7. The filename is provided as context but may not directly represent the title.

### Metadata Instruction Prompt For Lead Agency

Your task is to act as an expert environmental document analyst. From the provided Environmental Impact Statement (EIS) text (InputText), identify the federal or other agency responsible for leading the process. This information is typically found on the cover page, title page, executive summary, or introductory sections of the document.
First, provide your "reasoning" in a step-by-step manner. Explain how you located the lead agency in the text, considering potential variations in how the agency is named (e.g., a specific district or branch vs. the overarching department). Detail how you will map the identified agency to the standardized "Lead Agency - Lead Bureau" format required for the final output.
Based on your "reasoning", determine the "Lead Agency" and "Lead Bureau" that jointly represent the lead entity. Combine them into a single string using the exact format "Lead Agency - Lead Bureau" (e.g., "Department of Commerce - National Oceanic and Atmospheric Administration"). Your final "LeadAgency" output *must* be an exact match chosen from the following comprehensive list of standardized agency and bureau names. If no clear lead agency can be identified or mapped to the list, indicate that in your reasoning and select the most appropriate general category from the list if possible, otherwise explain why it cannot be found.
Here are some special cases: 1. If you find the lead agency to be "Federal Energy Regulatory Commission" (i.e., FERC) then you should return "Department of Energy - Department of Energy". 2. If you find the lead agency to be "Tennessee Valley Authority", then you should return "Other Independent Agencies - Tennessee Valley Authority". 3. If you find the lead agency to be "Department of Army, U.S. Army Corps of Engineers, xxx District", then you should return "Major Independent Agencies - Corps of Engineers–Civil Works".
List of acceptable Lead Agency & Bureau combinations: {agency_bureau_combination_list}.

### Metadata Instruction Prompt For Project Sponsor

Provide the name of responsible entity, organization, or person for this project. For externally proposed actions, this is typically the name of a company (or companies) that proposed and/or would carry out the proposed action. For actions proposed internally by a federal agency that is also leading the NEPA process, please indicate “None - action is sponsored by the lead agency”.

### Metadata Instruction Prompt For Project Type

Your task is to act as an expert environmental document analyst. Given the text of an Environmental Impact Statement (EIS) document, you need to select one or more project types from the following list that best match the description of the proposed action in the document: {project_type_list}.
Instructions: 1. Think of the project type as tags. Assign the tags that best match the description of the proposed action in the document. 2. If you find specific and general project types that both match the document text, then include both. For instance, "Surface Transportation - Railroads" is more specific than "Surface Transportation - Public Transportation". If they both match the document text, you can select both. 3. If none of the project types are a good fit with the proposed action, you may select "Other" as the project type. "Other" may be especially common for projects evaluated with categorical exclusions, which are highly variable. 4. The provided list of project types is designed primarily to capture the major categories of actions evaluated in environmental assessments and environmental impact statements.

### Metadata Instruction Prompt For Prepared By

Identify one or more agencies or entities (e.g., contractors) responsible for preparing a document based on the contents of the document. Categorical exclusion documents may list this information in the document heading or form fields, and are typically prepared by the lead agency. Environmental assessments and environmental impact statements often contain a list of preparers. Pay special attention to the contents of files with names similar to "List of Preparers".
It is not necessary to provide names of specific people, just their affiliated agencies or organizations. Provide full name of the agencies and organizations, do not use abbreviations. Note that, there might be a lead agnecy mentioned within the document, as well as, other agencies involved in the project. However, you should look for the agencies or organizations that prepared the document; i.e. they can be same or different.
The document can mention who the document was "prepared for". However, your task is to determine who the document was prepared by; i.e. they can be same or different.

## D Additional Description for Project Types

**Table 10.** Project types available on the NEPATECv2.

| **Project Type** | **Project Type Description** |
|---|---|
| Agriculture | Activities related to farming, ranching, and other agricultural practices, including crop production, livestock management, soil conservation, irrigation, pest control, and the conversion of land to agricultural uses. |
| Ecosystem Management and Restoration | Projects focused on maintaining, improving, or restoring the health, structure, and function of natural ecosystems. This includes activities like invasive species removal, wetland restoration, stream stabilization, and forest thinning for ecological benefit. |
| Habitat Conservation Plan | A planning document required by the Endangered Species Act that allows for incidental take of listed species in exchange for implementation of specific conservation measures that benefit the species. NEPA review is often required for the approval of an HCP. |
| Land Use or Forest Management Plan | Comprehensive plans guiding the use and management of land or forest resources within a specific area. This encompasses zoning regulations, resource allocation, silvicultural prescriptions, timber harvesting, recreational development, and fire management strategies. |
| Rangeland Management | Activities focused on the sustainable use and management of rangelands, including grazing management, vegetation manipulation, erosion control, and water resource development for livestock and wildlife. |

**Table 10 continued from previous page**

| Project Types | Project Type Description |
|---|---|
| Threatened and Endangered Species Management | Actions specifically designed to protect and recover species listed as threatened or endangered under the Endangered Species Act. This may include habitat restoration, captive breeding programs, population monitoring, and regulation of activities that could harm listed species. |
| Vegetation and Fuels Management | Projects aimed at controlling vegetation growth, reducing wildfire risk, and improving forest health. Includes prescribed burning, mechanical thinning, herbicide application, and other techniques to modify vegetation composition and structure. |
| Carbon Capture and Sequestration | Technologies and projects designed to capture carbon dioxide (CO2) emissions from industrial sources or directly from the atmosphere and store it permanently underground or in other long-term reservoirs. |
| Conventional Energy Production - Coal | The extraction, processing, and combustion of coal for electricity generation, industrial processes, and other energy needs. Includes coal mining, coal-fired power plants, and related infrastructure. |
| Conventional Energy Production - Land-based Oil & Gas | The exploration, drilling, extraction, processing, and transportation of oil and natural gas from onshore wells and facilities. Includes pipelines, storage tanks, and processing plants. |
| Conventional Energy Production - Nuclear | The generation of electricity through nuclear fission. Includes nuclear power plants, uranium mining and processing, and nuclear waste storage. |
| Conventional Energy Production - Offshore Oil and Gas | The exploration, drilling, extraction, processing, and transportation of oil and natural gas from offshore wells and facilities. Includes offshore platforms, pipelines, and onshore support facilities. |
| Conventional Energy Production - Other | Energy production from sources other than coal, oil, gas, and nuclear, which are still considered conventional based on widespread use and established technologies. This may include, for example, large-scale hydroelectric. |
| Conventional Energy Production - Rural Energy | Energy production projects using conventional sources (coal, oil & gas) with the purpose of supplying electricity, heating or other energy needs in rural communities. |
| Renewable Energy Production - Biomass | The generation of energy from organic matter, such as wood, crops, and agricultural waste. Includes biomass power plants, biofuel production facilities, and methane digesters. |
| Renewable Energy Production - Energy Storage | Projects that involve technologies designed to store electricity generated from renewable or other sources. Includes battery storage systems, pumped hydro storage, and compressed air energy storage. |
| Renewable Energy Production - Geothermal | The generation of energy from the Earth's internal heat. Includes geothermal power plants and direct-use geothermal applications. |
| Renewable Energy Production - Hydrokinetic | The generation of energy from the movement of water, such as tides, waves, and ocean currents. Includes hydrokinetic turbines and related infrastructure. |
| Renewable Energy Production - Hydropower | The generation of electricity from the force of flowing water. Includes dams, reservoirs, and hydroelectric power plants. This category typically refers to new or significantly altered hydroelectric projects. |

**Table 10 continued from previous page**

| Project Types | Project Type Description |
|---|---|
| Renewable Energy Production - Other | Renewable energy production from sources not specifically listed, such as algae-based biofuels or other novel renewable technologies. |
| Renewable Energy Production - Solar | The generation of electricity from sunlight using photovoltaic (PV) cells or concentrated solar power (CSP) systems. Includes solar farms, rooftop solar installations, and related infrastructure. |
| Renewable Energy Production - Wind, Offshore | The generation of electricity from wind using turbines located in marine or coastal environments. Includes offshore wind farms and related infrastructure. |
| Renewable Energy Production - Wind, Onshore | The generation of electricity from wind using turbines located on land. Includes wind farms and related infrastructure. |
| Laws, Policies, Regulations, and Guidance | The development, implementation, or modification of federal, state, or local laws, policies, regulations, or guidance documents that could have significant environmental impacts. This includes rulemakings, programmatic environmental assessments, and legislative proposals. |
| Land Development - Housing | The construction of new residential housing developments, including single-family homes, apartments, and condominiums. Includes site preparation, infrastructure development, and associated facilities. |
| Land Development - Other | The development of land for purposes other than housing, such as commercial buildings, industrial facilities, recreational areas, or infrastructure projects. |
| Land Development - Urban | Projects that entail the development or redevelopment of land within an already developed urban or suburban environment. |
| Public and Recreational Land Use | Projects pertaining to the development, maintenance, or improvement of public lands for recreational use, and related activities. |
| Manufacturing | The production of goods through mechanical, physical, or chemical processes. Includes factories, processing plants, and related infrastructure. |
| Mining - Metals | The extraction of metallic minerals from the earth. Includes open-pit mining, underground mining, and processing facilities. |
| Mining - Non-Metallic Minerals | The extraction of non-metallic minerals from the earth, such as sand, gravel, limestone, and gypsum. Includes quarries, mines, and processing facilities. |
| Mining - Marine Minerals | The extraction of minerals from the ocean floor or coastal areas. Includes dredging, underwater mining, and processing facilities. |
| Emergency and Disaster Response | Actions taken in response to natural disasters (e.g., hurricanes, earthquakes, wildfires) or other emergencies (e.g., oil spills, hazardous material releases) to protect public health and safety and mitigate environmental damage. |
| Military and Defense | Activities related to military operations, training, and infrastructure development, including military bases, weapons testing ranges, and defense-related research. |
| Research and Development | Scientific and technological research and development activities, including laboratory experiments, field studies, and pilot projects. |

**Table 10 continued from previous page**

| **Project Types** | **Project Type Description** |
|---|---|
| Routine Maintenance | Regular or recurring activities designed to keep existing infrastructure, facilities, or equipment in good working order. Includes road maintenance, building repairs, and equipment servicing. |
| Nuclear Technology | This sector includes research, development, deployment, and regulation of nuclear technologies beyond conventional energy production. This encompasses nuclear medicine, nuclear materials management, advanced reactor concepts, nuclear security, and other applications of nuclear science. |
| Other | A catch-all category for projects that do not fit neatly into any of the other categories. |
| Artificial Intelligence and Machine Learning | The development, deployment, and use of AI/ML technologies, including training models, developing algorithms, and deploying systems that could have environmental consequences. |
| Cybersecurity | Projects involving securing computer systems and networks from unauthorized access, use, disclosure, disruption, modification, or destruction. |
| Data Storage and Data Management | The construction and operation of data centers, data storage facilities, and related infrastructure for managing and storing large volumes of digital information. |
| High Performance Computing and Advanced Computer Hardware and Software | The development, deployment, and use of high-performance computing systems and advanced software applications for scientific research, engineering, and other computationally intensive tasks. |
| Quantum Information Science and Technology | Research, development, and deployment of technologies based on the principles of quantum mechanics. |
| Semiconductors | Design, manufacture, and testing of semiconductors and related microelectronic devices. |
| Aviation - Airports and Air Traffic | Construction, expansion, and operation of airports, air traffic control systems, and related infrastructure. |
| Aviation - Commercial Space | Activities related to the launch and operation of commercial spacecraft, including launch facilities, ground control stations, and satellite deployment. |
| Broadband | The deployment and expansion of high-speed internet infrastructure, including fiber optic cables, wireless towers, and satellite internet systems. |
| Electricity Transmission | The construction, operation, and maintenance of high-voltage power lines and substations for transmitting electricity over long distances. |
| Pipelines | The construction, operation, and maintenance of pipelines for transporting oil, natural gas, water, or other substances. |
| Ports and Waterways | Development, maintenance, and operation of ports, harbors, and navigable waterways, including dredging, channel construction, and the construction of docks and wharves. |
| Surface Transportation - Bridges | Construction, rehabilitation, or replacement of bridges for vehicular, pedestrian, or railroad traffic. |
| Surface Transportation - Other | Construction, reconstruction, or maintenance of highways, roads, and other surface transportation infrastructure (excluding bridges). |
| Surface Transportation - Public Transportation | Projects involving the development or improvement of public transportation systems, such as buses, light rail, subways, and commuter rail. |

**Table 10 continued from previous page**

| **Project Types** | **Project Type Description** |
|---|---|
| Surface Transportation - Railroads | Construction, operation, and maintenance of railroads, including track upgrades, rail yards, and passenger stations. |
| Utilities (electricity, gas, telecommunications) | Projects related to the provision of essential utilities, such as electricity, natural gas, telecommunications services, including distribution lines, pipelines, and communication towers. |
| Waste Management | Activities related to the collection, processing, treatment, and disposal of solid waste, hazardous waste, and wastewater. Includes landfills, incinerators, recycling facilities, and wastewater treatment plants. |
| Water Resources - Irrigation and Water Supply | Projects designed to provide water for irrigation, municipal water supply, or industrial use. Includes dams, reservoirs, canals, pipelines, and water treatment plants. |
| Water Resources - Other | Projects that involve water resource management but do not specifically fit within irrigation and water supply. This may include flood control projects, stream restoration projects, or wetland creation projects. |