Plan Overview

A Data Management Plan created using DMPonline

Title: Decoding Global Air Pollution Policies: A Three-Layer LLM-Based Framework for Constructing a Global Air Pollution Policy Database and Evaluating Policy Design

Creator:haiyu wu

Affiliation: University of Birmingham

Template: UoB short template

Project abstract:

Our Existing air-pollution-policy evaluation has advanced through the matching of structural emission breaks to policy events, yet no comparable global database exists for air pollution policies. This paper proposes a systematic, hallucination-robust framework for identifying and coding the universe of formal air pollution policies issued worldwide since 1970. We adopt a narrow operational definition that requires policies to directly target air quality, air pollutants, major emission sources, or air-quality governance mechanisms—excluding general climate, energy, or industrial policies unless they are explicitly operationalized for air pollution control. A three-layer corpus design is employed: a broad official environmental document universe (Layer 0), a high-recall candidate sub-universe (Layer 1), and a core global air pollution policy database (Layer 2) filtered through a two-round large language model (LLM) protocol with exact quote verification. The protocol integrates four positive inclusion criteria, two boundary remarks, and eight exclusion categories, alongside a country-specific issuer authority crosswalk and a dedicated classification for international agreements. All extractions are programmatically verified to prevent hallucination. Beyond binary inclusion, we construct a multi-dimensional policy design profile comprising stringency, specificity, implementation readiness, source coverage, multi-pollutant integration, air–carbon co-benefit potential, and a novel policy implementation cascade score (PICS) that captures the operationalization of parent policies through child regulations, standards, fiscal instruments, and enforcement mechanisms. The resulting database and design scores will underpin Chapter 2's empirical analysis of which air pollution policy designs trigger sustained clean-air and carbon co-reduction transitions.

ID: 209640

Start date: 04-01-2026

End date: 04-10-2026

Last modified: 19-08-2026

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Decoding Global Air Pollution Policies: A Three-Layer LLM-Based Framework for Constructing a Global Air Pollution Policy Database and Evaluating Policy Design

Data description

What types of data will be used or created?

This project will generate and utilize three primary types of data: (1) raw policy text files (PDFs, HTMLs, Word documents) collected via web scraping from official environmental databases (e.g., ECOLEX), ministries of environment websites across over 100 countries, and official government news platforms; (2) structured metadata for each policy document, including document identifiers, country, issuing authority, date, a multi-dimensional policy design profile (stringency, specificity, source coverage, air-carbon co-benefit potential), and a novel Policy Implementation Cascade Score (PICS) that captures operationalization through child regulations and enforcement mechanisms; and (3) Python scripts used for web scraping, data cleaning, LLM-based filtering, and scoring calculations, along with a country-specific issuer authority crosswalk and classification schemes for international agreements.

How will the data be structured and documented?

All raw policy texts will be stored in a hierarchical folder system organized by country, year, and source, with a master CSV/JSON index mapping each document to its metadata; all extracted design scores, PICS components, and LLM classification results will be maintained in structured CSV files and a SQLite relational database to track parent-child policy relationships; comprehensive documentation will include a data dictionary defining all scoring rubrics and inclusion/exclusion logic, detailed workflow logs of the two-round LLM verification process with exact quote verification records to ensure hallucination-robustness, and a top-level README file describing folder structures, naming conventions, and replication instructions, all stored alongside the data in open, non-proprietary formats.

Data storage and archiving

How will your data be stored and backed up?

All active project data—including raw policy files, metadata, and analytical outputs—will be stored on the University of Birmingham's Research Data Store (RDS), with access restricted to the research team via institutional authentication; the RDS is backed up on a daily basis with copies stored in separate physical locations from the live data, in compliance with the University's Backup and Retention Policy; additionally, all Python code and analytical scripts will be version-controlled using Git with periodic synchronization to the RDS, ensuring comprehensive data safety and recoverability throughout the project lifecycle.

Is any of the data of (ethically or commercially) sensitive nature? If so, how do you ensure the data are protected accordingly?

No, the data collected and generated in this project are not of an ethically or commercially sensitive nature, as all policy documents, official announcements, and metadata are derived exclusively from publicly available official government sources, international databases, and official news platforms, containing no personal data, trade secrets, or classified information; therefore, no specific anonymization or confidentiality procedures are required under UK GDPR or University data protection policies, although access to the live working dataset on the RDS will remain restricted to the core research team until formal publication to prevent premature dissemination or misinterpretation of unverified preliminary results.

Where will your data be archived in the long term?

Upon project completion and after the publication of core research papers, the final curated dataset will be archived in the University of Birmingham's eData repository (https://edata.bham.ac.uk/), which ensures long-term preservation, persistent identifiers (DOIs), and discoverability through major search engines; an embargo period of up to 12 months may be applied to protect academic priority and allow for the completion of all related publications, following which the dataset will be made openly accessible under a permissive license (e.g., CC BY), ensuring that other researchers can freely use and build upon the global air pollution policy database.

Data sharing

Which data will you share, and under which conditions? How will you make the data available to others?

Following project completion and publication, the final curated core global air pollution policy database (Layer 2) including all policy metadata, design profiles, and PICS scores, alongside the complete coding framework, data dictionary, and Python scripts for data processing and scoring, will be deposited in the University of Birmingham's eData repository and shared under an open CC BY 4.0 license, with a Data Access Statement included in all related publications linking directly to the dataset's DOI; a 12-month embargo may be placed to coincide with thesis submission and journal publication timelines, and while the full raw text corpus of Layer 0 and Layer 1 documents will not be shared directly due to data volume and third-party copyright constraints, a complete list of document URLs and DOIs will be provided so users can retrieve original texts independently.