RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 10 (2026) 1–15

A Curated Multi-Source Corpus of Climate Finance Literature, 1990–2024: Multilingual Retrieval and Institutional Reports

Minh Ha-Duong (Corresponding Author)
CIRED/CNRS, Paris, France, minh.ha-duong@cnrs.fr

Received 26 March 2026, revised 29 July 2026, accepted 30 July 2026

Abstract

This data paper presents a curated, multilingual corpus of 33,344 works on climate finance published between 1990 and 2024. The dataset is assembled from 8 complementary sources that combine academic databases with a selected layer of institutional reports and key documents (1.4% of the corpus), adding institutional vocabulary and negotiation records absent from academic indexes. A multilingual retrieval strategy based on an eight-language keyword taxonomy is used to capture relevant works across linguistic contexts, while a reproducible pipeline integrates deduplication, metadata harmonisation, and quality filtering. English accounts for 93.8% of the refined works. The corpus includes a citation network derived from Crossref and OpenAlex, as well as embeddings pre-computed by a multilingual sentence-transformer.

Keywords

climate finance – bibliometric corpus – multilingual – sentence-transformer embeddings – scientometrics – history of economic thought

Related Data Set

"A Curated Multi-Source Corpus of Climate Finance Literature, 1990–2024" DOI 10.5281/zenodo.19236130 in repository "Zenodo".

1. Introduction

Climate finance, that is North–South financial flows directed at mitigating and adapting to climate change, has become a centrepiece of climate change conferences. Yet the scholarship discussing it remains bibliographically fragmented. It includes both academic publications, dispersed across databases with different coverage profiles, and institutional publications from organisations such as the UNFCCC (diplomatic negotiation records), the OECD (bilateral and multilateral finance guidelines and statistics), or the Climate Policy Initiative (think tank reports) not indexed in standard academic databases.

Published with license by RDJ | **DOI:** 10.52024/fn48qf66
 | ISSN: 2452-3666 (online)

The literature is also multilingual: while English dominates, a monolingual approach risks missing entirely the multicultural and Southern perspectives.

Pioneering bibliometric work has begun to map this terrain. Carè and Weber (2023) survey climate finance using Scopus; Shang and Jin (2023) analyse climate finance from Web of Science; Maria et al. (2023) map green finance with network analysis. These studies rely on a single database each, search in English only, and they publish only the query, not the reusable dataset.

The present work addresses these limitations: a climate finance corpus built from eight complementary sources, with a documented quality-filtering pipeline, searched in eight languages, and embedded with a sentence-transformer trained to place texts in any of these languages in one semantic space. Because "climate finance" is a contested category whose boundaries are themselves an object of study, we err on the side of inclusion and publish a much larger corpus than the three studies discussed above. A query-level probe bounds the overlap: replicating each study's published query against OpenAlex (matched by DOI and year-constrained title; probe of 2026-07-22), the corpus contains 89.3% of the works retrieved for Carè and Weber's (2023) query and 91.0% for Shang and Jin (2023)'s. Those studies searched Scopus and Web of Science, so these shares also absorb OpenAlex coverage gaps and query-translation drift: they approximate the overlap with the records those authors analysed. The lower 40.1% for the query by Maria et al. (2023) reflects scope as much as retrieval, since it spans green finance at large — a broader field than this dataset covers by design.

The dataset was constructed for a history-of-economic-thought study of how climate finance emerged as an economic aggregate [Ha-Duong, under review], but its scope makes it reusable for topic modelling, citation network analysis, bibliometric mapping, and case studies of scholarship published in languages other than English.

2. Method

2.1. Sources: Discovery and Ingestion

The corpus assembles academic literature and institutional documents from 8 sources with complementary coverage profiles (Table 1). Five are automated or semi-automated (reproducible from the scripts and an internet connection); three require manual export from restricted platforms.

The search strategy uses a four-tier keyword taxonomy reflecting the evolving vocabulary of climate finance: core terms in eight languages (Tier 1: English, French, German, Spanish, Portuguese, Arabic, Chinese, Japanese), English institutional and diplomatic vocabulary (Tier 2), and climate-adjacent terms retained only when the abstract mentions two (Tier 3) or three (Tier 4) of

Table 1: Sources, Automation Level, and Coverage Scope

Source	Automation	Coverage
OpenAlex (Priem et al., 2022)	Automated (free API)	Primary academic source (indexes Crossref, DOAJ, PubMed, and more): tiered keyword search
ISTEX	Automated (public metadata API; fulltext restricted to French research institutions)	CNRS national text-mining platform (multi-publisher)
Institutional reports	Hybrid (curated seed + World Bank API)	17 selected reports (e.g., Buchner et al., 2013) + World Bank repository grey-literature harvest
Teaching canon	AI-assisted (scraping + LLM extraction)	Syllabus readings from university courses worldwide
bibCNRS	Hand-harvested (CNRS credentials)	CNRS portal aggregating non-English news & discourse (Gale, Wanfang, NewsBank) in FR, ZH, JA, DE
SciSpace	AI-collected, hand-exported	SciSpace systematic review tool exports
UNFCCC	AI-collected	Negotiation documents
OECD	Hand-harvested	Statistical guidelines

four concept groups (climate, finance, development, environment). Every term is issued against OpenAlex's default.search field, which matches title, abstract, and indexed fulltext. The taxonomy was informed by keyword mining of 2,644 core papers (cited ≥ 50 times).

The full protocol per source, with query fields, term counts by tier, and language coverage, is deposited as tab_retrieval_protocol.csv, whose companion.md also enumerates the curated institutional-reports seed list in full. The taxonomy, concept groups, title blacklist, and curated seed lists themselves ship as commented YAML in the deposit (openalex_queries.yaml, corpus_filter.yaml, grey_sources.yaml).

All API queries are bounded to publication years 1990–2024; the lower bound reflects the start of UNFCCC negotiations, before which "climate finance" had no institutional referent, and the upper bound stops short of the harvest date because indexing lags make the databases incomplete for the most recent years. The harvest pool is append-only, so 9.8% of refined works nonetheless carry a year outside the window, mostly later (Section 2.4). OpenAlex indexes economics working papers from RePEc, NBER, MPRA, and other repositories, so no separate harvest is needed. Per-source record counts, before and after refining, appear in Table 3; the bibCNRS aggregates capture non-English public discourse absent from academic databases. The teaching canon was built by scraping university course syllabi worldwide and extracting bibliographic

references through a combination of DOI regex matching and LLM-assisted extraction, followed by fuzzy deduplication and Crossref enrichment (details in the companion technical report). For UNFCCC and OECD documents available in multiple languages, we index only the English version.

The UNFCCC and OECD layers enter through a curated inclusion rule rather than a keyword query. A document qualifies if it is an official document produced by or addressed to the COP or the OECD DAC that defines, accounts for, or governs climate finance, including party and negotiating-bloc submissions, together with the negotiation record around those decisions: session summaries such as the IISD Earth Negotiations Bulletin, and statements delivered at COPs. The classes were fixed empirically rather than by an a priori list of institutions, by crossing a companion history-of-thought study's bibliography against the first corpus version and keeping the classes of documents it cited but the corpus lacked.

2.2. Deduplication and Refinement

The merge pipeline combines records from all sources through two deduplication passes: DOI-based (normalised, lowercased), which removes 833 records, and title+year matching for records lacking DOIs, which removes a further 159. Enrichment can backfill a DOI onto a record that entered without one, revealing duplicates the first two passes could not match; a third DOI pass after enrichment removes them (counted in Table 2). When duplicates are found, the maximum citation count is retained, and metadata follows a source-priority order.

Boolean from_* columns (one per source) track which databases contributed each record, enabling provenance analysis. The source_count field records multi-source agreement: of the 33,344 refined works, 738 (2.2%) appear in multiple sources.

Table 2: Corpus Construction Ledger

Stage	In	Removed	Out
DOI deduplication at merge	44,174	833	43,341
Records without a title dropped	43,341	3	43,338
Title and year deduplication at merge	43,338	159	43,179
Quality filtering, protection criteria applied	43,179	9,436	33,743
Duplicate-DOI records dropped after enrichment	33,743	399	33,344

Each stage starts where the previous one ended; In less Removed equals Out on every row. The first In is the pooled source-catalog records, the last Out the refined corpus.

Each work in the combined deduplicated list is then examined and annotated on a handful of facets. Three flags require no external data: (1) missing metadata (no title), (2) absent abstract combined with an irrelevant title (lacking domain-specific terms in any of the target languages), and (3) title blacklist matches (noise terms such as "blockchain" or "deep learning" without climate-finance context). Two flags depend on enrichment outputs: (4) citation isolation (pre-2020 works neither cited by nor citing any other corpus work), and (5) cross-encoder relevance scoring against the query "climate policy and financial mechanisms", which flags works scoring below 0.002. The deposited code also implements a sixth, semantic-outlier flag; it was not active in this build — its calibration on a language-imbalanced corpus is unsettled — and the deposited corpus carries no such column.

A separate grouping step detects near-duplicate content: coordinated multi-journal publications where the same text appears under different DOIs. Detection uses two passes: clustering on the first 200 characters of the normalised abstract, then clustering on normalised titles, a title group qualifying only when at least 50% of its members share an abstract prefix with another member. Both passes require five members, and matched works receive a shared near_duplicate_group identifier. These records are kept, not removed: they are real publications.

Four further facets confer protected status, marking works whose relevance is established independently of the flags: citation count ≥ 50, presence in 2 or more sources, within-corpus citations, or appearance in teaching syllabi.

The refine step draws the consequence: a *refined* work is one that is unflagged or protected; the rest, along with the few records dated before 1960, are removed from the working corpus but retained in the deposit with their flags. Every inclusion/exclusion decision is recorded in the audit trail. Table 2 traces the corpus from the pooled source records to the 33,344 refined works, one row per stage. Of the 11,244 works flagged (counts are non-exclusive: a single work may trigger multiple flags), citation isolation (2,677) and cross-encoder relevance scoring (5,840) are the most common, followed by absent abstracts with irrelevant titles (2,397), missing metadata (753), and title blacklist matches (423); 1,808 of them meet a protection criterion and survive the filter.

2.3. Quality and Completeness

This section examines the corpus on three fronts: what each source contributes and how complete its metadata is (Table 3), how accurate and complete the citation graph is against Crossref and how far its reference lists reach beyond the corpus, and how well the relevance scores that drive filtering agree with human judgment.

Table 3 summarises corpus composition and metadata completeness per source; a work discovered by several sources counts in each of their rows.

Table 3: Corpus Sources

Source	Raw	Refined	Unique	%non-EN	%DOI	%Abstract	%Refs
OpenAlex	41,138	31,544	30,815	6%	76%	88%	61%
ISTEX	748	635	443	2%	100%	97%	96%
bibCNRS	233	219	200	8%	10%	5%	1%
SciSpace	663	618	238	1%	83%	90%	66%
Institutional reports	281	210	103	0%	95%	97%	32%
Teaching canon	622	618	549	9%	100%	31%	55%
UNFCCC key documents	232	230	225	0%	0%	94%	0%
OECD DAC key documents	35	33	33	0%	12%	94%	6%
TOTAL	**43,179**	**33,344**	**32,606**	**6%**	**76%**	**87%**	**61%**

Raw: records with from_* provenance flag before filtering (a record in multiple sources is counted once per source). Refined: after six-flag quality filtering. Unique: found only in that source (source_count = 1). The **TOTAL** row is the deduplicated union of works, not the column sum: 748 raw records and 738 refined works carry more than one provenance flag, which puts the source rows 773 and 763 memberships above their totals (25 and 25 of them carry three). %non-EN: share of non-English works. %DOI, %Abstract, %Refs: metadata completeness among refined records.

OpenAlex contributes 94.6% of refined works; the seven smaller sources increase the corpus's non-English, institutional, and pedagogical coverage. The "Unique" column counts the works only that source contributed (source_count = 1), and it shows the sources are complementary rather than redundant: between 39% (SciSpace) and 100% (OECD DAC key documents) of each smaller source's refined works appear in no other source. %non-EN is the share of non-English works, %DOI the share with a valid DOI, %Abstract the share with an abstract longer than 50 characters, and %Refs the share with at least one reference in the citation graph.

Citation-graph accuracy was audited on a simple random sample of 300 links drawn, with a fixed seed, from the links whose citing and cited works both carry a DOI. A sampled link counts as confirmed when the cited work's DOI appears in the reference list the citing work's publisher deposited at Crossref: 97.0% of the sample is confirmed (95% CI [94.4%, 98.4%]). The 9 unconfirmed links appear in our data but not in that deposited list, which is itself incomplete for books, chapters, and older articles, so absence there does not establish that a link is wrong. Tested in the other direction, the corpus captures 98.3% of the reference DOIs Crossref does hold (6,016 of 6,119, 95% CI [98.0%, 98.6%]).

Both checks measure agreement with Crossref, not ground truth, and the confirmation is partly circular for the links that were themselves harvested from Crossref or OpenAlex deposits; they bound link accuracy and capture, not the share of true citations the graph misses. Restricting to links whose citing work belongs to the refined corpus yields 1,087,209 outgoing citation pairs (the cited work may lie outside the corpus); 80% of DOI-bearing corpus works appear as a citing source; among the core papers (cited ≥ 50 times), that share rises to 96%. The stricter subset where both endpoints are corpus works is the graph of Figure 1: it connects 13,112 works, 39% of the corpus.

The number of references each work contributes to the citation graph varies by orders of magnitude, so it doubles as a screening variable. Among the

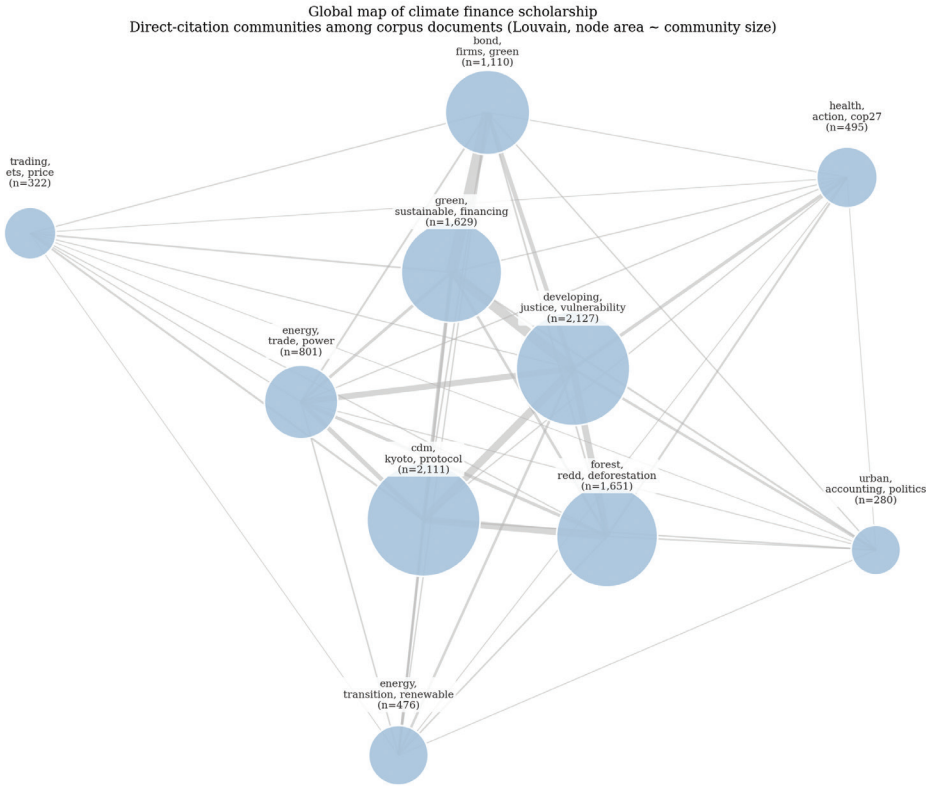

Figure 1: Communities in the Connected Subgraph of the Corpus Citation Network. Each circle is a community: area proportional to the number of works, edge width to the number of citations between communities, label from the community's top TF-IDF terms. Underneath, each work is a node and an edge links it to each corpus work it cites (DOI-matched); Louvain detection (fixed seed, modularity 0.62) groups the works. The map draws the 10 communities holding at least 2% of connected works, which together cover 84% of the 13,112 works with at least one such link (39% of the corpus). The force-directed layout is indicative — distances carry no measurement.

25,411 DOI-bearing works, the median is 29 references (mean 42.8, 95th percentile 126, maximum 1,536); 5,159 works (20.3%) contribute none, a gap concentrated in books, reviews, editorials, and grey literature, while counts in the hundreds mark review articles.

The reference lists reach far beyond the corpus, and following them outward is reverse snowballing: harvesting the works the corpus cites rather than the works the queries retrieve. The deposited citation table sizes that outer ring: the refined works cite 280,990 distinct DOIs, of which 97.1% are not corpus members; 113 of the absent works are cited fifty times or more from within the corpus, and the most-cited of them are classics of neighbouring literatures — environmental economics, panel econometrics, climate science, green-bond finance. The corpus contains the works its search protocol retrieved, not their references; the deposited citations.csv names the cited works by DOI, title, and year, so the snowball ring is reachable from the deposit without a new harvest.

Relevance filtering (flag 5) scores each work with a cross-encoder reranker (BAAI/bge-reranker-v2-m3) against the query "climate policy and financial mechanisms" and flags works scoring below a fixed threshold. Validation used a blinded sample of 100 works stratified by score quintile, graded in randomised order: the human-relevant share rose monotonically across quintiles (10%, 15%, 20%, 60%, 80%). Because the quintiles are score-ordered, these rates alone fix the discrimination of the scores at AUC = 0.818. At the deployed threshold the grading recorded 81% accuracy (precision 74%, recall 76%), and halving or doubling the threshold reclassifies roughly 10% of the calibration sample. The deposit ships this evidence: tab_reranker_validation.csv (the per-quintile rates, from which the AUC recomputes), reranker_hitl_stratified.csv (the validation sample; its per-work grades were not retained, so the per-quintile rates are the surviving record), reranker_hitl_review.csv (a second, threshold-zone sample with per-work grades), and reranker_calibration.csv (the weak-label calibration set).

2.4. Limitations and Biases

Several biases qualify reuse. The corpus is strongest on the academic pool and weakest on policy discourse and media, partially captured through the institutional-reports and bibCNRS sources. Non-English coverage remains limited (Table 4): English accounts for 93.8% of refined works despite targeting eight languages: indexing systems amplify English's established position as the language of scientific communication; the classified non-English layer counts 2,061 works, led by Portuguese, Spanish, German, and French, and a further 11 carry no language tag — coverage that supports case studies rather than balanced comparison.

Filtering compounds the language imbalance: 41.8% of non-English works are removed, against 19.9% of English ones, and the cross-encoder flag

Table 4: Language Distribution in the Refined Corpus

Language	Code	Works	Share (%)
English	en	31,272	93.8
Portuguese	pt	418	1.3
Spanish	es	308	0.9
German	de	295	0.9
French	fr	256	0.8
Other (38 languages)	—	784	2.4
Unclassified	—	11	0.0
Total		33,344	100.0

accounts for 58.9% of those non-English removals. Metadata coverage does not explain the gap, since non-English works have the higher abstract coverage. The scorer itself does, at least in part: rescoring the same non-English works against a same-language translation of the query lifts the share crossing the deployed threshold by 11.5 percentage points, so the score partly reflects query language rather than topical relevance. The per-stratum removal ablation ships as tab_filter_ablation.csv; users needing non-English coverage can rebuild from the deposited flags under their own filter policy.

The corpus is not strictly confined to its 1990–2024 search window: 3,278 refined works (9.8%) carry publication years outside it, 3,238 after 2024 and 40 before 1990. They entered because the year bounds were added to the API queries partway through collection (March 2026) and the harvest pool is append-only: records gathered by the earlier, unbounded queries stay in the pool, and no later pipeline stage re-applies the window. Some pre-1990 years are metadata errors — several works carry the placeholder year 1970 — and others are older classics from the teaching canon.

Citation counts reflect OpenAlex data as of 2026-07-24 and favour older, English-language works; the teaching canon is biased toward institutions with publicly accessible syllabi, and its LLM extraction step has not been independently audited. The keyword taxonomy is retrospective: mined from the vocabulary the field settled on, it embeds the negotiation calendar in the harvest itself, so dating the field's turns from this corpus alone partly re-measures the taxonomy.

The two curated key-document layers carry almost no DOIs and no reference lists (Table 3), so they contribute text and embeddings but stay outside the citation network: an analysis run on the citation graph sees only the academic corpus.

The institutional-reports source is scoped to international institutions (the UNFCCC system, the OECD, the World Bank, transnational think tanks) and deliberately excludes documents from national institutions: government reports, economic analysis councils, central bank communications. Only its World Bank component is a repository harvest; the rest is a selection, so the layer samples rather than surveys institutional writing. The boundary follows the research object (climate finance as an economic aggregate for international economic governance, constituted in the UNFCCC and OECD arenas) and tractability. We acknowledge that the exclusion costs multi-cultural coverage, since these institutions are dominated by OECD countries. The corpus could be complemented with regional databases such as CNKI for Chinese journals, where domestic scholarship on green finance is extensive, SciELO for Latin American journals, where Brazilian scholarship on green and climate finance is active, or Garuda for Indonesian journals, for example. The data harvesting pipeline is also available, and can accept additional sources to extend the corpus without reprocessing existing ones.

OpenAlex offers the widest coverage of any open scholarly database and indexes working papers, but its metadata quality is uneven (Priem et al., 2022). It supplies 94.6% of refined works; the documented flaws and the pipeline responses follow.

Keyword queries return false positives. The flag-based filter (Section 2.2) is the response; its cross-encoder relevance scoring was human-validated at AUC = 0.818 (Section 2.3).

Some duplicate records survive. Two passes deduplicate at merge (DOI-based, then title+year), and near_duplicate_group annotates same-content works published under distinct DOIs. Some working paper / published version pairs still slip through: works sharing an exact normalised title and first-author surname but published up to five years apart form 344 candidate pairs (1.0% of works), about half of them genuine version pairs; fuzzy title matching raises the upper bound to 1,329 works (4.0%), a tier dominated by distinct report editions. The passes also over-merge occasionally: 38 DOI groups join records with unrelated titles, and the title+year key merges 18 generically titled empty-year works into 7 groups. The direction of the residual bias is known: surviving version pairs double-count citations, inflating co-citation and community-size statistics. Author metadata is itself noisy across sources, so author-normalised deduplication is left to a future release. The audit script and its output table (tab_dedup_error_estimates.csv) ship with the pipeline code.

The abstract field concentrates two flaws. Abstracts are often missing: enrichment reconstructs them from OpenAlex inverted indices and ISTEX fulltext, and residual absences trigger flag 2. Stored abstracts are sometimes contaminated, holding boilerplate (paywall notices, repository metadata) or the article's own text: boilerplate detection excludes stub abstracts from the embeddings, and oversize detection excludes full-text fields.

Language codes are often missing or non-standard: enrichment backfills them from the OpenAlex API, then infers what remains with langdetect from title and abstract, and language_provenance marks each tag's origin. The inference covers 1,331 refined works (4.0%).

Citation assignments can be incorrect; the audit reported in Section 2.3 confirmed 97.0% of sampled links, and the remainder are references our data carries that the citing work's Crossref deposit does not list.

Reference lists arrive empty or aberrant; per-document reference counts (Section 2.3) are the screening variable, and GROBID parses the unstructured reference strings Crossref and OpenAlex leave unresolved, recovering title, first author, and year for fuzzy matching against corpus works (corpus_parse_citations_grobid.py in the deposited code). The audit above verifies the links we have, not the links we miss: with a fifth of DOI-bearing works contributing no outgoing edge, network statistics should be checked for sensitivity to their exclusion.

We do not hold the full texts of the corpus's academic works, so the citation graph relies on indexed metadata and deposited reference strings, and no-DOI works (24%) are undercounted — a limitation this release does not mitigate.

3. Data Description

The Zenodo deposit (DOI: 10.5281/zenodo.19236130, temporal coverage 1990–2024) is a reproducibility archive in three parts: the pipeline source code (code/: scripts, configuration, and DVC definitions), the raw data inputs (data/inputs/: the 8 per-source catalogs as harvested, before merge and deduplication), and the final data products of this paper (data/products/: the v2 corpus files described below). All CSV files use UTF-8 encoding with standard comma-separated format; identifiers are opaque string keys.

climate_finance_corpus.csv (43,179 rows). The complete, unfiltered corpus: one deduplicated work per row, every column named and described in Table 5. A machine-readable descriptor, datapackage.json, ships beside the data: a Frictionless Table Schema declaring each column's storage type, allowed values, range, and share of empty cells, validated against the CSV at packaging time. The refined subset (33,344 works) starts from df[~df['is_flagged'] | df['is_protected']], then drops 399 duplicate DOIs surfaced by post-filter enrichment and 2 pre-1960 records; the deposited corpus_audit.csv records every keep, removal, and deduplication with its reason, so the refined subset reconstructs exactly from the deposit. Abstracts are not included in the deposited file due to publisher redistribution restrictions; they can be retrieved programmatically via the DOI or OpenAlex identifier, and the abstract_status column records each abstract's fate: original, reconstructed from an inverted index or fulltext, LLM-summarised, oversized, or missing.

Table 5: Variables of climate_finance_corpus.csv, in Four Groups: Record Identity, Bibliographic Metadata, Provenance Flags, Curation Metadata

Variable	Description
source	Primary source catalog for the record's metadata
source_id	Identifier in the primary source (e.g. OpenAlex work ID)
doi	Digital Object Identifier, when available
title	Title of the work
first_author	First author name
all_authors	Full author list, separator-joined
year	Publication year
journal	Publication venue (journal, publisher, or repository)
language	Language code (ISO 639-1), detected and normalised
keywords	Keywords, semicolon-separated
categories	Subject categories / concepts from the source catalog
cited_by_count	Citation count (OpenAlex, as of the collection date)
affiliations	Author affiliations, when available
from_openalex	Provenance flag: found in OpenAlex
from_istex	Provenance flag: found in ISTEX
from_bibcnrs	Provenance flag: found in bibCNRS
from_scispace	Provenance flag: found via SciSpace
from_grey	Provenance flag: institutional reports (Section 2.1)
from_teaching	Provenance flag: teaching canon (syllabi)
from_unfccc	Provenance flag: curated UNFCCC key document
from_oecd	Provenance flag: curated OECD key document
abstract_provenance	Provenance of the abstract, for curated key documents only
keywords_provenance	Provenance of the keywords, for curated key documents only
language_provenance	How the language code was obtained (Section 2.4)
source_count	Number of sources that contributed the record
abstract_status	Fate of the undistributed abstract (Section 3)
near_duplicate_group	Group id of near-identical content under several DOIs
is_flagged	Any quality flag raised (refined-subset rule: Section 3)
flag_reason	Comma-separated raised quality flags; empty when unflagged
is_protected	Protection from removal (key papers kept despite flags)
protection_reason	Why the work is protected (Section 2.2)

Generated from the deposit column contract (scripts/_deposit_variables.py); storage types, allowed values, ranges and measured missingness are in the deposited datapackage.json.

embeddings.npz. Compressed NumPy archive (keys: vectors as float32 array, keys as string array of identifiers) containing L2-normalised (unit-length) 1024-dimensional vectors for 38,736 of the 43,179 works, computed by the multilingual sentence-transformer *BAAI/bge-m3*. Works without a title are excluded; works with missing or boilerplate abstracts (paywall stubs, repository metadata) are embedded from title and keywords only. Row *i* of vectors corresponds to keys[i]. Flagged works are included to support alternative filtering strategies. Each vector encodes the concatenation of title, abstract (when available and non-boilerplate), and keywords within the model's 8192-token context window (no truncation for typical abstracts). The model maps texts in English, French, Chinese, Japanese, and German into a shared semantic space, which is designed to group works by topic rather than by language. Vectors still carry the language-specific confound: we did not subtract it. The vectors and the reranker relevance scores were computed from the abstracts as fetched at build time; a re-fetch returns the index's current state, so recomputing them reproduces our values only up to index drift.

citations.csv. Citation network (1,375,310 rows). Columns: source_doi, ref_doi (both normalised, lowercased DOIs). Each row records a (citing, cited) pair where the citing work belongs to the full deposited corpus of 43,179 works, flagged records included — restricting the citing side to the refined subset leaves 1,087,209 rows; the cited work may or may not belong to the corpus. For corpus-internal citations only, join on the corpus identifiers. The graph operates on DOI pairs, so no-DOI works are under-represented on both sides, as citers and as cited targets (Section 2.4).

Source catalogs. Per-source catalogs (openalex_works.csv, istex_works.csv, etc.), shipped under data/inputs/, record each database's contribution as harvested, before merge and deduplication. The corpus does not include a dedicated Open Access or fulltext URL column; users can recover OA links via the OpenAlex API using the source_id field, or construct ISTEX fulltext URLs from ISTEX identifiers in the source catalogs.

The dataset is released under CC BY 4.0. The five automated or semi-automated sources can be re-harvested with an internet connection; the three restricted sources (bibCNRS, SciSpace, OECD) have their raw exports included in the Zenodo deposit.

4. Data Overview

This section summarises the corpus from two angles: its growth in time, reported in the text, and the structure of its citations, mapped in Figure 1.

Annual publication volume grows from a handful of works in the early 1990s (the search window is 1990–2024; the earliest works date from 1992) to a take-off with a structural break in the trend of log annual counts at 2015 (Chow test

of one log-linear trend against separate intercept and slope on each side, classical errors: $F = 10$, $p = 0.0004$), output growing 17% a year since. The phrase "climate finance" is nearly absent from titles and abstracts before 2009, then spreads through the literature. The annual-volume figure appears in the companion study (Ha-Duong, under review).

Citation coverage — the share of corpus works that appear as a citing source in the citation network — rises across the three periods: 40% for works published before 2007, 47% for 2007–2014, and 69% for post-2015 works. Works without a DOI cannot be matched as citing sources, so the undercount concentrates in early and grey-literature works.

The second angle is citation structure. Figure 1 maps the connected subgraph of the corpus-internal citation network — the 13,112 works (39% of the corpus) holding at least one DOI-matched link to another corpus work: Louvain community detection on that graph yields 10 major communities covering 84% of the connected works, and the communities correspond to recognisable research programmes, from carbon markets to green bonds.

The companion paper [Ha-Duong, under review] gives an alternative overview of the corpus contents through a semantic clustering of the deposited embeddings; its themes resemble the citation communities without matching them, since works that share a vocabulary do not necessarily cite each other.

5. Concluding Remarks

This corpus assembles climate finance scholarship into a single multilingual bibliographic object, pairing the breadth of OpenAlex with French archives (ISTEX), non-English discourse (bibCNRS), pedagogical use (teaching canon), AI-assisted discovery (SciSpace), and a selected layer of institutional reports and UNFCCC and OECD key documents that carries the negotiation record and the institutions' own vocabulary into an otherwise academic corpus. Its provenance and author metadata support networks of institutions and co-authors, its embeddings feed topic models and content classification, and its eight-language retrieval adds a non-English layer that the English-only queries of the prior mappings cannot reach; the architecture accepts new sources and a longer temporal window, and a future iteration would use OpenAlex's bulk download rather than keyword queries. By making its boundaries explicit rather than implicit, the corpus turns the definition of "climate finance" from a hidden assumption into an object of analysis.

Data and Code Availability

The corpus data files and pipeline source code are deposited in Zenodo under CC BY 4.0 at 10.5281/zenodo.19236130. Key software dependencies: Python

≥ 3.10, pandas, sentence-transformers (with PyTorch), and DVC. Suggested citation:

Ha-Duong, M. (2026). A Curated Multi-Source Corpus of Climate Finance Literature, 1990–2024 [Data set]. Zenodo. 10.5281/zenodo.19236130.

Acknowledgements

This work was supported by CNRS (Centre National de la Recherche Scientifique) through the author's permanent research position at CIRED (Centre International de Recherche sur l'Environnement et le Développement).

AI Disclosure

Claude (Anthropic) was used as a research assistant throughout: extracting bibliographic references from university course syllabi for the teaching canon, implementing and testing the corpus pipeline, and drafting revisions under the author's direction. The author designed the corpus and its quality-filtering policy, arbitrated every editorial decision, and verified the results; extracted references were validated against Crossref metadata. The author takes full responsibility for the content.

References

Buchner, B., Stadelmann, M., Wilkinson, J., Mazza, F., Rosenberg, A., & Abramskiehn, D. (2013). *The global landscape of climate finance 2013*. Climate Policy Initiative. https://www.climatepolicyinitiative.org/publication/global-landscape-of-climate-finance-2013/

Carè, R., & Weber, O. (2023). How much finance is in climate finance? A bibliometric review, critiques, and future research directions. *Research in International Business and Finance*, *64*, 101886. https://doi.org/10.1016/j.ribaf.2023.101886

Maria, M. R., Ballini, R., & Souza, R. F. (2023). Evolution of green finance: A bibliometric analysis through complex networks and machine learning. *Sustainability*, *15*(2), 967. https://doi.org/10.3390/su15020967

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. https://doi.org/10.48550/arXiv.2205.01833

Shang, Q., & Jin, X. (2023). A bibliometric analysis on climate finance: Current status and future directions. *Environmental Science and Pollution Research*, *30*(57), 119711–119732. https://doi.org/10.1007/s11356-023-31006-5