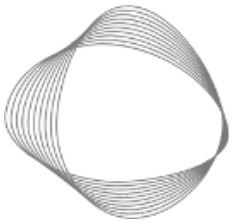
Climate TRACE Ownership Information: Source & Company-Level Ownership Methodology



Lee Gans[1,6], Gavin McCormick[1,6], Ishan Saraswat[1,6], Gabriela Volpato[1,6], Krsna Raniga[1,6], Christy Lewis[1,6], Jeremy Freeman[1,6], Aaron Davitt[1,6], Lauren Schmeisser[2,6], Nils Jenson[2,6], Joseph Fallurin[2,6], Andrew Isabirye[3,6], Ashank Sinha[3,6], Matthew Jolley[4,6], and Kerri De Sousa[5,6]

1) WattTime, 2) RMI, 3) TransitionZero, 4) Hypervine Limited, 5) OceanMind, 6) Climate TRACE

Overview

This document describes how source-level and company-level ownership data were produced for the following sectors: Cement, Chemicals, Pulp and Paper, Oil and Gas Production and Transport, Oil and Gas Refining, Petrochemicals, Aluminum, Iron, Bauxite, Copper and Coal Mining, International and Domestic Aviation, Shipping, Electricity Generation, and Steel.

1. Introduction

Independent reports quantifying the emissions generated by private companies and state-owned enterprises (for-profit enterprises owned by governments) are few and far between. For instance, in 2016, Science.org published a report estimating that just 90 companies generated close to two thirds of industrial greenhouse gas (GHG) emissions for the year 2013 (Heede, 2014; Starr, 2016). The underlying datasets for this report, which estimated the companies' yearly emissions based on historical production numbers, took nearly a decade to collect. With regard to state-owned enterprises (SOEs), in 2022, Columbia's SIPA Center for Global Energy Policy published a report estimating that 300 SOEs emitted 7.49 gigatons of CO2 equivalent in Scope 1 emissions during 2017 (Clark & Benoit, 2022). This report included the caveat that "the true scale of SOE-related emissions is likely to be substantially higher, particularly when accounting for national oil companies and iron and steel manufacturers that do not currently report their emissions". As these two examples demonstrate, aggregating and analyzing the datasets necessary to produce independent emissions estimates for companies and SOEs is incredibly challenging and time-consuming. As a result, such reports are often out of date by the time they are published, and/or limited in the scope of their coverage. Currently, the most frequent and up-to-date emissions estimates for private companies and SOEs are self-reported inventories. There is a plethora of incentives that can potentially bias self-reporting or discourage doing so at all – such as regulatory consequences, public outcry, and loss of investors. The companies whose GHG emissions are disproportionately large are also disproportionately unlikely to self-report their emissions unless they are required to do so. Therefore, having up-to-date, independent

emissions estimates with extensive, worldwide coverage for private companies and SOEs is an essential component of the actionable information that investors, policymakers, and activists need in the fight to reduce carbon emissions.

For 2021, Climate TRACE generated the first ever independent, openly accessible source and company-level emissions database – providing ownership information for 4,342 private companies and SOEs, with comprehensive coverage of emissions-generating activity in 8 sectors. For 2022, we have expanded our ownership database to include 13,159 private companies and SOEs, located in 234 countries and administrative regions, and with comprehensive coverage for 17 sectors. Our 2022 database continues to be inclusive of sectors in which it has historically been the most difficult to track ownership and estimate emissions: electricity generation, steel, and fossil fuels.

In last year's data release, Climate TRACE provided source-level ownership for 27.2% of global estimated emissions. For our release this year, we are providing source-level ownership data for 38% of global estimated emissions. For 2021, we made ownership data openly available for download directly from our website. This data included the names of direct owners and ultimate parents, as well as the percent share of ownership attributable to direct owners. This year, direct downloads from our website include the names of all identified owners, the percent shares of all direct owners and ultimate parents, and citations for all available primary sources for each ownership claim (either directly in the download, or available via a link in the download to a page on the GEM wiki). Even more information about owners is available for users upon request. To whatever extent feasible, Climate TRACE collected data quantifying intermediate levels of ownership (links in the chain of ownership in between the direct owner and ultimate parent). Furthermore, over half of the owners in our database are additionally linked to widely-used, international standards for identifying unique legal entities, such as legal entity identifiers (LEIs), PermIDs, or stock ticker symbols. This renders it easier than ever to cross-reference (with high confidence) our emissions-by-owner database with other datasets that institutions use for supply-chain management. For the year 2022, the subset of owners that are linked to international ID systems was responsible for an estimated 26% of global emissions. Even more ownership metadata –such as official company website URLs (6% of owners) and registered addresses (62% of owners) –is also available on request.

2. Materials

Identification of company-level ownership occurred in two stages. First, datasets containing source-level owners were aggregated into sector-specific datasets. For this analysis, a 'source-level owner' is defined as the company whose name is attached to a single individual source in a dataset. Once source-level ownership datasets were aggregated, research was conducted to identify the source-level owner's ultimate parent company, subsidiaries, joint

ventures, and sibling companies. In particular, identifying *higher-level* rungs in the ownership chain was prioritized, with lower levels being included only when they incidentally turned up during research on higher levels. Often, the lowest-level identified owner is not actually the lowest-level subsidiary in the ownership chain for a source, but merely the lowest level that was reported in a primary dataset. Since the goal of the ownership database is to attribute responsibility for emissions to institutions, it was a priority to follow chains all the way up, but not necessarily all the way down. In the download packages, the column "percent_company_datasource" contains citations that link an emissions source to direct owners, and the column "percent_parent_datasource" contains citations that link the direct owner to its ultimate parent. To generate company-level emissions estimates, sources that share an ultimate parent were aggregated together and their emissions were summed. Emissions (and, where applicable, production) estimates for each ultimate parent include the sum of all estimates for associated sources, in proportion to the ultimate parent's percentage of ownership for each one. Download packages contain raw data that can be re-aggregated in this way to reproduce company-level estimates. Pre-aggregated company level estimates can be generated on request.

2.1 Source-level Ownership Datasets by Sector & Methodology

Assembling the Climate TRACE ownership database was a coalition-wide effort with contributions from many members. Information on direct owners (as in, the lowest-level identified owner in the chain of ownership) was provided by RMI, TransitionZero, Hypervine, OceanMind, WattTime, and Global Energy Monitor (GEM). Meanwhile, information connecting these lowest-level, direct owners up to their highest-level ultimate parents (e.g. corporations, investment firms, and governments) was generated by GEM and WattTime. A diverse set of methodologies and data sources were used both to identify direct owners and map them up to ultimate parents. This document provides a broad overview of the methodologies and data sources used for the entire Climate TRACE ownership dataset, with particular specificity and analysis pertaining to the automated mapping method employed by WattTime specifically. For a more detailed discussion of Global Energy Monitor's methods, see Global Energy Monitor's methodology.

For this section, the following source-level ownership terms were used in each sector:

- 'Real property' is defined as a parcel of land along with any permanent fixtures attached to it.
- 'Property' is defined as a tangible object that is not land, nor affixed to it.
- A 'business concern' is a private legal entity formed for the purpose of engaging in commercial activity. A 'state-owned enterprise' (SOE) is defined as a for-profit business owned by a government.
- A 'government agency' is a non-profit entity owned by a government.
- 'Percent financial interest' is defined as the proportion of the monetary value of an asset or of an organization to which the holder of the interest is legally entitled

2.1.1 Global Energy Monitor (GEM) Ownership-Mapping Sectors

Global Energy Monitor (GEM) specializes in providing the highest standard of verification for each individual ownership claim contributed to the Climate TRACE ownership database. GEM employs teams of sector specialists, language experts, and trained ownership researchers to generate both direct and ultimate parent ownership data through desk research. Because GEM derives ownership data from vetted, official primary sources wherever possible, and manually verifies each claim, users can be confident that GEM ownership data is high quality in terms of accuracy and recency. For all sectors in this section, GEM aggregated both direct ownership and mapped direct owners up to ultimate parents.

Electricity Generation. Sources were defined at the level of individual power plants. Ownership was defined in terms of percent financial interest in the asset as a piece of real property, a business concern, state-owned enterprise, or government agency. Ownership data for 4492 plants was derived from the GEM Wiki's Global Coal Plant Tracker (GCPT) and Global Gas Plant Tracker (GGPT) (GEM, 2022). Ownership for an additional 527 plants was reported using data from the U.S. Energy Information Administration (EIA, 2022) and aggregated by WattTime. In some cases, there were different owners for specific units within each source. In these cases, the unit level ownership data was aggregated to the asset level by summing the ownership of each unit in the asset weighted by its capacity. See the Climate TRACE electricity methodology for more information on plant capacity.

Steel. Sources were defined at the level of individual steel manufacturing facilities. Ownership was defined in terms of percent financial interest in the asset as a piece of real property, a business concern, state-owned enterprise, or government agency. The source-level ownership data source for steel is the Global Energy Monitor's (GEM) Global Steel Plant Tracker (GSPT). GSPT provides facility-level data for all steel plants that produce at least 0.5 million metric tons of crude steel per annum (GEM, 2022).

Coal Mining. Sources were defined at the level of individual mines as reported by the GEM Wiki Global Coal Mine Tracker (GCMT) and CoalSwarm project (GEM, 2022). Ownership was defined in terms of percent financial interest in the asset's coal reserves (GEM, 2021).

Oil and Gas Refining – Teapot Refineries. 'Teapot refineries' are a subset of small, privately-owned oil and gas refineries in China. Sources were defined at the level of individual facilities engaged in oil and gas refining. Ownership was defined in terms of percent financial interest in the asset as a piece of real property, a business concern, state-owned enterprise, or government agency.

2.1.2 WattTime Ownership-Mapping Sectors

For all sectors in this section, WattTime mapped direct ownership datasets up to ultimate parents. Mapping was performed using WattTime's novel, automated ownership mapping algorithm that collects webscraped and API-derived ownership information from large, freely available entity relationship datasets. The algorithm matches records in relationship databases based on international entity-id systems and linguistic pattern matching. Manual review and desk research were performed for low-confidence matches and unmatched entities leftover after automated mapping. Manual, quantitative validation was performed on a subset of the results of the mapping algorithm to ensure accuracy and high-quality. Direct ownership for these sectors was generated from the methods and data aggregators listed below.

Oil and Gas Production and Transport. Sources were defined at the level of oil and gas fields, or macro-geological formations. Source boundaries were derived from shapefiles included in an industry dataset licensed from Rystad Energy. Percent ownership for each source was defined based on the owner's percent financial interest in the total oil and gas production for the source. Rystad provides owner name and ownership percentage for most Climate TRACE sources. In some cases, ownership data was available only for subfields within sources. In these cases, the ownership percentage was aggregated based on the proportion of total production attributable to the owner across all subfields vs. total production for all subfields within the asset. See Climate TRACE oil and gas methodology for more information about how production and assets were defined. Direct ownership was aggregated and sources were defined by RMI.

Oil and Gas Refining. Sources were defined at the level of individual facilities engaged in oil and gas refining. Ownership was defined in terms of percent financial interest in the source as a piece of real property, a business concern, state-owned enterprise, or government agency. Ownership for refineries was derived from desk research using company websites, government sources, and news articles. Direct ownership was aggregated and sources were defined by RMI.

Petrochemicals. Sources were defined at the level of individual facilities engaged in the production of petrochemicals. Ownership was defined in terms of percent financial interest in the source as a piece of real property, a business concern, state-owned enterprise, or government agency. Ownership for petrochemical facilities was derived from desk research using company websites, government sources, and news articles. Direct ownership was aggregated and sources were defined by RMI.

Cement. Sources were defined at the level of individual cement manufacturing facilities. Ownership was defined in terms of percent financial interest in the source as a piece of real property, a business concern, state-owned enterprise, or government agency. The Spatial Finance Initiative (SFI) provides a Global Database of Cement Production Assets (GDCPA) that includes

facility-level ownership data. The database is part of the GeoAsset Project developed by SFI, Oxford Sustainable Finance Programme, Satellite Applications Catapult, and The Alan Turing Institute (McCarten *et al*., 2021). Climate TRACE provides emissions estimates for clinker-producing plants identified by the GDCPA (see the Climate TRACE cement methodology for discussion of cement production methods and emissions model). Ownership data was available from the GDCPA for 1,504 of these plants (Armstrong *et al*., 2021; Global Cement Directory, 2021). Direct ownership was aggregated and sources were defined by TransitionZero.

Chemicals. Sources were defined at the level of individual facilities engaged in the production of industrial chemicals. Ownership was defined in terms of percent financial interest in the source as a piece of real property, a business concern, state-owned enterprise, or government agency. Ownership for chemical facilities was derived from desk research using company websites, government sources, and news articles. Direct ownership was aggregated and sources were defined by TransitionZero.

Aluminum. Sources were defined at the level of individual facilities engaged in the production of aluminum. Ownership was defined in terms of percent financial interest in the source as a piece of real property, a business concern, state-owned enterprise, or government agency. Direct ownership for aluminum facilities was derived from industry publication, Light Metal Age (Light Metal Age, 2022). Direct ownership was aggregated and sources were defined by TransitionZero.

Pulp and paper. Sources were defined at the level of individual facilities engaged in the production of paper products. Ownership was defined in terms of percent financial interest in the source as a piece of real property, a business concern, state-owned enterprise, or government agency. Ownership for pulp and paper facilities was derived from desk research using company websites, government sources, and news articles. Direct ownership was aggregated and sources were defined by TransitionZero.

International and Domestic Aviation. Sources were defined at the level of airports. Ownership was defined in terms of the percentage of the source's emissions attributable to each airline or 'operator' at the airport. For each operator, Climate TRACE calculated the proportion between the emissions attributable to the operator's flights vs. all flights at the asset overall. Half of the total emissions for each individual flight were attributed to the flight's airports of departure and arrival, respectively. Emissions for international and domestic flights were estimated separately. The Official Airline Guides (OAG) Historical Flight Status Data identifies flights by airline and airport for all domestic and international flights –including passenger, commercial, private, and general aviation. Direct ownership was aggregated and sources were defined by WattTime.

Shipping. Ownership was reported at the level of companies. Ownership was defined in terms of percent financial interest of the ultimate parent in all sources (individual vessels) as pieces of real property. Ownership for shipping assets was provided by Lloyd's List (Lloyd List Intelligence, 2022). Direct ownership was aggregated and sources were defined by OceanMind.

Copper, Iron, & Bauxite Mining. Sources were defined at the level of individual mines as reported by Hypervine. Ownership was defined in terms of percent financial interest in the mine's mineral reserves. Ownership for chemical facilities was derived from desk research using company websites, government sources, and news articles. Direct ownership was aggregated and sources were defined by Hypervine.

2.2 Ownership Data Coverage by Sector & Data Aggregator

Table 1 summarizes the coverage of ownership information for the 15 sectors in which Climate TRACE provides source-level emissions estimates and ownership data, and 2 sectors (international and domestic shipping) where we provide source-level emissions and company-level ownership data.

Table 1 Asset and Emissions Coverage for Ownership Data by Sector

Sector	# Sources Emissions Estimated	# Sources Ownership Identified	# Sources w/ Ownership %	# of Countries	% Emissions Sources	% Emissions Sector	% Emissions Global	Ultimate Parents	Direct Owners
Electricity Generation	7883	5019	64%	143	96%	83%	18.34%	GEM	GEM & WattTime
Oil & Gas Production & Transport	676	617	91%	86	59%	57%	5.72%	WattTime	RMI
Steel	873	839	96%	72	77%	73%	2.82%	GEM	GEM
Coal Mining	3161	2987	94%	66	100%	92%	2.27%	GEM	GEM
Cement	2259	1504	67%	114	76%	54%	2.19%	WattTime	TransitionZero
Oil and Gas Refining	683	572	84%	103	99%	91%	1.51%	WattTime	RMI
+ Teapot Refineries		101	15%	1		8%	0.14%	GEM	GEM
International & Domestic Aviation	4891	4891	100%	231	100%	100%	1.33%	WattTime	WattTime
Chemicals	477	477	100%	65	96%	96%	0.94%	WattTime	TransitionZero
Aluminum	229	229	100%	44	100%	100%	0.60%	WattTime	TransitionZero
International & Domestic Shipping	-	-	-	-	-	-	0.58%	WattTime	OceanMind
Petrochemicals	116	116	100%	21	52%	52%	0.25%	WattTime	RMI
Copper Mining	555	496	89%	52	100%	94%	0.13%	WattTime	Hypervine
Iron Mining	536	496	93%	41	98%	98%	0.12%	WattTime	Hypervine
Pulp and Paper	383	265	69%	40	100%	73%	0.10%	WattTime	TransitionZero
Bauxite Mining	175	141	81%	22	98%	96%	0.01%	WattTime	Hypervine

In the table above, the "# of countries" column shows how many countries contain at least one source with an identified owner in the sector. The "% Emissions" columns show the percentage of 2022 global emissions derived from sources with identified owners at three levels- "% Emissions Sources": emissions from all identified sources, "% Emissions Sector": emissions from the entire sector (source-level + remaining country-level emissions), and "% Emissions Global": total sector emissions that Climate TRACE has ownership data for as a percentage of global estimated emissions. The columns "Ultimate Parents" and "Direct Owners" display which coalition member provided each level of ownership data for the final Climate TRACE ownership database. Due to licensing restrictions, international and domestic shipping include only the "% Emissions Global" column because ownership in those sectors is not reported at the level of sources (individual ships). Instead, this data is reported at the level of companies.

3. WattTime's Automated Entity Relationship Mapping Algorithm

3.1 Automated Mass Mapping Datasets

GLEIF. The Global Legal Entity Identifier Foundation (GLEIF) is an open, global, industry database of legal entities that is curated by McKinsey & Company. Legal Entity Identifiers (LEIs) are international business credentials that companies can apply to receive. The typical reason companies apply for an LEI is to gain access to global financial markets outside their country of origin. As such, the GLEIF database is limited only to such companies that have applied for and received an LEI credential, resulting in a dataset that primarily includes large, international conglomerates and their mid-level international shell companies. GLEIF corporate relationship data was downloaded in bulk from the GLEIF website. An example of GLEIF relationship data for Mitsubishi Group, a parent company of several steel and coal mining assets, can be found here. GLEIF entities from bulk data were pre-processed and cleaned for the purposes of matching (as described in section 4.1 below). LEI relationships were identified at the level of ultimate parents, direct parents, and child entities. GLEIF is especially useful for identifying large corporations headquartered in the global south.

PermID. The PermID database is administered by financial market data provider, LSEG. On their website, PermID is described as "a machine readable identifier that provides a unique reference for data item(s). Unlike most identifiers, PermID provides comprehensive identification across a wide variety of entity types including organizations, instruments, funds, issuers and people. PermID never changes and is unambiguous, making it ideal as a reference identifier… intended to enable interoperability." PermID provides users with "a set of descriptive metadata for each entity to facilitate disambiguation to PermID with or without explicit mapping." PermID data was accessed using the OpenPermID API with the OpenPermID package for Python. PermID is the primary provider of registered address and official website url data for the Climate TRACE ownership database.

Wikidata. Wikidata is a crowdsourced database that contains information from Wikipedia that has been transformed by users into an analyzable format. Using the SPARQL query service, users can generate relevant datasets and download them en masse. For applicable sectors, SPARQL queries were constructed to return a dataset of all relevant entities that were instances of the following: business, company, conglomerate, enterprise, holding company, corporation, multinational corporation, government agency, public company, state-owned enterprise, and township-level division. In order to ensure the entities returned by the query were relevant to the sector, additional properties were specified. For example, queries were constructed such that the entity needed to either be involved in a specific industry (ex. oil refining) or make a specific product (ex. cement). An example SPARQL query that returns the Wikidata dataset for the steel sector can be found here. A subset of this data was then used to perform corporation mapping.

First, entities whose Wikidata entries included alternate company names and/or relationships such as 'owner of', 'has subsidiary', 'owned by', 'parent organization', and 'partnership with' were added to the mapping dataset for the sector. Additional entities and relationships were added to each dataset through web scraping the actual text of the Wikipedia page links returned by the SPARQL query. Web Scraping was performed using the Python package for Selenium Webdriver with Chrome. The scraper returned in-text lists of subsidiaries. SPARQL queries return Wikidata results from pages in every language available on Wikipedia, with relevant data types transliterated into English. Wikidata entries from non-English language pages were included in the dataset. Meanwhile, scraping of page text was limited to English language pages only. Because of its non-English language coverage and ability to query information about state-owned entities, Wikidata was an especially useful source for international companies, especially in China. Because both LEIs and PermIDs are Wikidata properties, all entities identified through those datasets were queried based on those properties to gather additional information. Although Wikidata is crowdsourced, datasets include citation links so that it is possible to track down the source for each data point. All primary sources linked in Wikidata articles were extracted from SPARQL query results, and are listed (along with the Wiki Page they were found on) as sources in the data source table.

OpenCorporates. OpenCorporates is the largest open, crowdsourced database of business entities in the world. Although the data is crowdsourced, it is also curated such that sources only include public official information. OpenCorporates provides many types of current and historical information, such as company officers, registered addresses, agents, statements of control, subsidiaries, branches, and similarly named companies. An example of an OpenCorporates entry for Berkshire Hathaway, the parent company for three of the top 500 emitting power plants in the Climate TRACE dataset, can be found here.

OpenCorporates has a partnership with OpenRefine, a tool created by Code for Science & Society. OpenRefine is designed for working with messy, publicly available data and mapping it onto other open data sources. The OpenCorporates Reconciliation service available on OpenRefine was used to generate matches for asset-level owners based on the owner's name and the country of operation. The reconciliation service provides a scoring system for matches, and initial matches with scores below 50 were manually reviewed for accuracy. Once the initial matches were generated, the list was input to the OpenCorporates API using the Ropencorporate package for R. The API was queried to return a list of parents, subsidiaries, branches, and controlled companies for each match where such information was available. Corporate

relationships were identified through the explicitly listed data points returned through the API, and also by identifying entities that shared CEO's, presidents, managers, and registered addresses in common. Because OpenCorporates is derived from local, official business registries, it is an especially useful source for low-to-mid level shell companies. However, its coverage for companies, especially ultimate parent companies, outside the global north is limited.

SEC Filings. Securities and Exchange Commission (SEC) Filings were used as a mass automated mapping dataset for oil and gas sectors only. For the top 50 global oil and gas producers, PDFs were downloaded from the SEC database. Specifically, these PDFs contained an appendix titled 'Subsidiaries of the Registrant' from 10-K and 20-K annual reports for 2021. An example from ExxonMobil can be found here. The text from these PDFs was parsed using the PyPDF2 package for Python and aggregated into a single dataset. This dataset was aggregated exclusively for oil and gas sectors because international oil and gas companies are especially likely to be massive conglomerates or SOE's that do business in the USA. In 2021, the USA was the second largest net oil importer in the world, and the largest oil producer (WorldsTopExports, EIA). SEC filings also provide information about percentage ownership for subsidiaries and joint ventures. SEC filings data was pre-processed and mapped as described in Section 4.1 below.

3.2 Validation Datasets

For three sectors, it was possible to compare company-level estimates for production with company-level production estimates derived from industry or government datasets.

Steel. For last year's ownership database, company-level Climate TRACE estimates of production were compared with self-reported production estimates from the WorldSteel Association for the top 114 steel producers in 2021 (WorldSteel Association, 2021). This year, automated mapping for steel was unnecessary because GEM provided a full, manually mapped dataset for steel. However, this provided a good opportunity to ensure that automated mapping produced results that were equivalent to results produced entirely through human vision. Although GEM's definition for ultimate parents differs slightly from the definition used for the database last year and from how producers self-define themselves to the WorldSteel Association, an analysis comparing estimates for production based on manually mapped, automatically mapped, and self-reported steel production estimates was performed as a validation for this year's ownership database. However, the published data available for download for steel this year, reflects the manual mapping performed by GEM and GEM's definition of ultimate parents (see Global Energy Monitor's ownership methodology for more details).

Oil and Gas Production and Transport. For 2022, percent ownership for oil and gas production from all assets in Texas was compared with equivalent estimates based on data from the Texas Railroad Commission (RRC). Texas assets alone account for 8.7% of global emissions for the sector. Two datasets were aggregated from the RRC. First, annual production data paired with lease ids were scraped from the Oil & Gas Production Data Query system. Then, a dataset matching lease ids to owner names was parsed from PDFs (see Oil & Gas Lease Name Index). These datasets were then joined by lease id. The owner names were pre-processed and mapped as described in Section 4.1.

Cement. Company-level Climate TRACE estimates of production were compared with self-reported production estimates from the Global Cement Directory for the top 50 cement producers in 2022 (Global Cement Directory, 2022).

4. Methods

4.1 Pre-Processing

Prior to mapping corporate networks, owner names from asset-level datasets, Wikidata, PermID, and GLEIF were consolidated, cleaned, and transformed into several standard formats for the purposes of matching. Relationship databases such as OpenCorporates and Mergr.com were queried using LEIs and PermID's to identify high-confidence record matches and based on unique identifiers from these systems. Then, further mapping and consolidation was performed using a combination of algorithmically generated groupings of potentially related entities that were verified using manual desk research where necessary.

Consolidation. Raw ownership data from all source-level datasets frequently contained several name variants for individual entities (ex. 'Mitsui & Co. Ltd', 'Mitsui and Company', 'Mitsui &Co '). Any one of these variants could potentially produce a match from one of the corporate mapping datasets. For this reason, instead of transforming all of these into a single standardized format for matching (which could eliminate useful variations), the first step was to identify which name variants within the dataset likely refer to the same entity and assign them a key. That way, if a match is returned for one variant, it can instantly be applied to the others that share its key.

Variants were identified using OpenRefine's Cluster and Edit functions. The Cluster and Edit functions are approximate string matching algorithms designed to find text strings that contain the same content, even if they are spelled differently. OpenRefine's clustering methods include: Fingerprint, N-gram Fingerprint, Metaphone3, Cologne-Phonetic, Daitch-Mokotoff, Beider-Morse, Levenshtein Distance, and Prediction by Partial Matching (PPM). See OpenRefine's technical reference for an in-depth explanation of each method. Clusters identified

through these functions were manually reviewed for accuracy. Each cluster was then assigned a key. The key was based on the most common name variant in the cluster and was constructed using a customized version of OpenRefine's default fingerprint method. The steps in OpenRefine's default fingerprint method are shown in Table 2 (OpenRefine User Guide, 2022).

Table 2 Processing Steps to Produce OpenRefine's Default Fingerprint

Processing Step	Example String 1	Example String 2
Remove leading and trailing whitespace	Companhia Siderúrgica Nacional (CSN)	Mitsui & Co. Ltd
Change characters to lowercase	companhia siderúrgica nacional (csn)	mitsui & co. ltd.
Remove all punctuation and control characters	companhia siderúrgica nacional csn	mitsui co ltd
Normalize extended western characters to ASCII	companhia siderurgica nacional csn	mitsui co ltd
Split the string into whitespace-separated tokens	['companhia'] ['siderurgica'] ['nacional'] ['csn']	['mitsui'] ['co'] ['ltd']
Sort tokens alphabetically and remove duplicates	['companhia'] ['csn'] ['nacional'] ['siderurgica']	['co'] ['ltd'] ['mitsui']
Join tokens back together into fingerprint string	companhia csn nacional siderurgica	co ltd mitsui

For this analysis, OpenRefine's default fingerprint method had to be modified to increase accuracy during clustering. For each sector, there were specific words that were common across the dataset but were only rarely relevant to identifying unique entities. These irrelevant portions of the string were often longer than the relevant ones. Often, they were shared between unrelated companies, producing false positives. Alternatively, these irrelevant strings varied too much between genuinely related entities, producing false negatives. Table 3 shows the default fingerprints for a cluster of similarly named owners from the steel assets dataset, alongside the portions of these strings that are relevant for clustering vs. the irrelevant portions that tend to create errors.

Table 3 Default vs. Customized OpenRefine Fingerprints

Original Owner Name	Default Fingerprint	Relevant	Irrelevant
Anyang Iron & Steel Co., Ltd.	anyang co iron ltd steel	anyang	co iron ltd steel
ANYANG IRON AND STEEL GROUP CO., LTD.	and anyang co group ltd steel	anyang	and co group ltd steel
Anyang Xinpu Steel Co., Ltd.	anyang co ltd steel xinpu	anyang xinpu	co ltd steel
Anyang Iron and Steel Co Ltd	and anyang co iron ltd steel	anyang	and co iron ltd steel
Angang Steel Co., Ltd.	angang co ltd steel	angang	co ltd steel

Customized fingerprints included only the relevant parts of the default fingerprint string (Table 3). In order to identify irrelevant parts of strings, OpenRefine's word facet was used to review the 100 most common words in the dataset. Then, irrelevant sector-specific words were identified and removed (ex. 'steel', 'cement', 'plant', 'power', 'exploration'). Additionally, a list of common country-specific company abbreviations and terms for business entities were removed for the relevant countries automatically (ex. 'Ltd', 'PT', 'Companhia'), as well as the word 'and'. Exceptions for the word 'and' as well as '&' characters were made in cases where those strings appeared between consecutive single consonants (ex. 'S & N Drilling'). In such cases, 'S & N' was treated as a single string. With these terms removed, the clustering functions were applied, and the resulting clusters were assigned a modified fingerprint as a key.

In the example in Table 3, three unique keys – 'anyang', 'anyang xinpu', and 'angang' –were produced from the original owner names. Although they were similar, they were not consolidated down further. 'Anyang' could be a company name, or a province name, or both –and 'angang' could be a typo for 'anyang', or it could be a separate entity entirely. As it turned out, matches uncovered during the mapping process confirmed that all three of these keys are distinct shell companies that fall under the same corporate parent: Ansteel Group. However, in other cases (ex. 'Shaanxi' vs. 'Shanxi') slight variants referred to entirely different entities. Hence, the clustering process errs on the side of keeping assumptions conservative.

4.2 Reconciliation among Corporation Mapping Datasets

After consolidation, the original owner names from the datasets were transformed into several formats for matching. Table 4 shows an example of each format for the owner 'Pangang Group Jiangyou Changcheng Special Steel Co., Ltd.'.

Table 4 Examples of Owner Name Formatted for Matching

Variant	Format Type	Formatted Name
1	Original Owner Name	Pangang Group Jiangyou Changcheng Special Steel Co., Ltd.
2	Company terms removed at end of string only	Pangang Group Jiangyou Changcheng Special Steel
3	Key	changcheng jiangyou pangang
4	First two words, lowercase, no punctuation	pangang group
5	Key, sector specific words intact	changcheng jiangyou pangang special steel

For reconciliation with Wikidata and GLEIF, matching was performed using all variants (1-5) – like for like – between the mapping dataset and the asset-level dataset. Information from Wikidata and GLEIF were also cross-referenced with each other, so that ultimate parent groupings would include matches from both. The most similar match for the longest string in a

cluster was considered the best match for the group. For reconciliation with OpenCorporates, it is not practical to download bulk data and match transformed names. So, the standard reconciliation process was used to match only variants 1 and 2 in Table 4. Then, these matches were expanded into groupings based on results from the API, and only the novel owner names generated by the API were matched and cross-referenced using the other three variants (3-5). OpenCorporates reconciliation was attempted first while specifying the country of operations. For unmatched entities, it was attempted again without specifying a region.

2.2 Desk Research for Unmatched Owners

In cases where the mapping process uncovered ambiguous (equally similar and conflicting) matches, or if it failed to return a match, desk research was conducted. Desk research included custom, manual searches for government and industry news sources, and within the mapping datasets. It also included searching the following sources: SEC filings, official company websites (ex. Baosteel), and industry news and data sources (ex. mergr). Companies that still remained unmatched or lacked additional nodes in their corporate network at the completion of mapping were considered to be their own ultimate parents.

2.3 Criteria for Assigning Direct Parents and Final Owner Groupings

Once the data were fully mapped, it was necessary to establish criteria for reporting ultimate parents for the purposes of inclusion in the Climate TRACE database. The current analysis was designed to produce reliable data en masse that would enable reasonably accurate and independent emissions estimates at the company-level. It was also designed to produce accessible, summary information for users of the Climate TRACE website. In some cases, some of the nuance and complexity of the corporate relationships data that was returned had to be simplified down in a standardized way. In other cases, the datasets available did not specify components that relate directly to calculating emissions (i.e., percentage ownership). To handle this issue, general and sector-specific criteria were established to simplify complex and incomplete data. These criteria include the following:

General. For assets in which over 50% of the owners were individual persons, the owner is listed as ‘Unknown’ or ‘persons’ –even if their identity was listed in the original source-level dataset. As a policy, Climate TRACE does not publish the identities of private individuals.

Steel. According to their 2020-21 report, the WorldSteel Association calculates the list of top steel producers as follows: “In case of more than 50% ownership, 100% of the subsidiary's tonnage is included, unless specified otherwise. In cases of 30%-50% ownership, pro-rata tonnage is included. Unless otherwise specified in the declaration, less than 30% ownership is considered a minority and therefore, not included.” The same accounting scheme was used for the purposes of validating Climate TRACE production estimates and assessing company-level emissions. For owners with more than one direct parent where the parent’s percent interest was

unknown, whichever direct parent was largest (produced the most steel) was chosen.

Oil and Gas Production and Transport. In the United States, there are far more private oil and gas production enterprises than is typical in other countries. For instance, in the Rystad dataset, there are 434 unique owner names associated with assets outside the United States, while there are 492 unique owner names associated with US domestic assets. This proportion is not in keeping with the proportion of oil production between the US and other countries, nor with the proportion of coverage for the US vs. internationally in the Rystad dataset. The main reasons for this are that most other countries have more nationalized oil and gas assets, and the United States has a uniquely venture-capital friendly system for oil and gas production. Private equity firms often provide the seed money for multiple, small, independent, private oil and gas producers. As a result, SEC filings may not list these companies as subsidiaries, and they are less likely to appear in the mapping datasets. For companies operating in the United States, if an oil and gas company website or industry news source listed a private equity firm as an owner's sole financier, the firm was listed as the company's parent. Although parent is not always the correct technical term to describe such a relationship (i.e., a seed investor, etc.), this choice was made to increase the utility of the dataset in facilitating the reduction of emissions in the sector. This approach provides the advantage of quantifying emissions in a way that puts the spotlight on the actions of large institutional actors as opposed to small-to-mid-level businesses.

For SOE's, data on which entities are SOE's and the governments they are associated with can be provided as additional data on request. Because the Climate TRACE ownership database provides enhanced insight into the roles of both private equity firms and world governments in the oil and gas sectors, Climate TRACE's dataset is well-suited to provide more actionable information for investors, policymakers, and activists alike.

3. Selected Results

3.1 Steel

For steel, Climate TRACE's model estimates production as part of its emissions model. Importantly, while Climate TRACE uses WorldSteel's production estimates at the country level as a baseline to generate these production estimates, Climate TRACE does not use company-level WorldSteel estimates as part of this process. Company-level estimates are self-reported by companies with assets that are often spread between many different countries. If the mapping methodology used in this analysis is accurate, then production estimates from Climate TRACE should match the company's self-reported production numbers, up to the percentage of the company's assets that are included in the Climate TRACE dataset.

Figure 1 shows a comparison of production rankings from WorldSteel and Climate TRACE for 2021. Production is measured in tonnes of crude steel, where white represents the 1st largest producer and the darkest purple represents the 30th largest producer, according to WorldSteel.

The middle column shows WorldSteel's rankings, which serve as the baseline for comparison for Climate TRACE's rankings. The left column shows Climate TRACE's rankings for each of the top 30 steel producers in the WorldSteel dataset before mapping. These rankings reflect Climate TRACE's total estimated crude steel production for all assets whose original, unmapped owner names contained the names of the WorldSteel's top 30 steel producers. The right column shows Climate TRACE's ranking for WorldSteel's top 30 after mapping. These rankings reflect Climate TRACE's total estimated production for all assets whose original owner names were clustered and mapped to their appropriate corporate parent. After mapping, there is a complete correspondence between WorldSteel's rankings (based on self-reporting from the companies) and Climate TRACE's rankings (based on modeling asset-level production and mapping of crowdsourced and primary source ownership datasets).

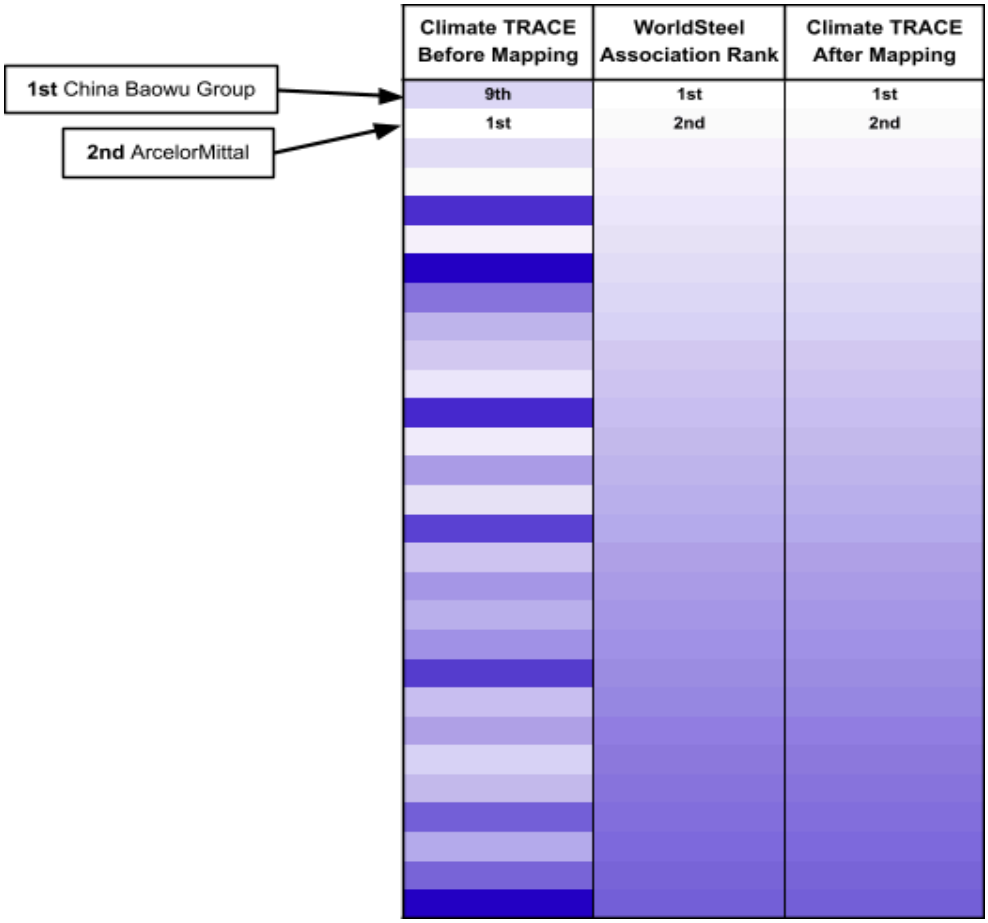


Figure 1 Ranking of 2021 Production Estimates from Climate TRACE Before and After Mapping. In the figure, China Baowu Group was ranked 1st for steel produced followed by ArcelorMittal based on self-reports to the WorldSteel Association (boxes with arrows on the left). However, prior to mapping each company's subsidiaries, Climate TRACE placed ArcelorMittal 1st and China Baowu Group 9th (Left column in the figure). After subsidiaries were included through mapping, Climate TRACE's rankings matched those of WorldSteel for all top 30 companies in the dataset.

As shown in Figure 1, prior to mapping, Climate TRACE's company-level production rankings reflected more noise than signal. Pre-mapping rankings were often determined by the company's naming conventions for its subsidiaries, rather than actual differences in production. For example, WorldSteel identified China Baowu Group as the #1 steel producer in 2021 by far, producing 40 million tonnes (34%) more steel than its nearest competitor, ArcelorMittal. The difference in production between these two companies was the largest difference between two consecutively ranked companies in the entire WorldSteel dataset, in both absolute and percentage terms. Nevertheless, in the unmapped Climate TRACE dataset, China Baowu Group was ranked 9th, and ArcelorMittal, 1st. This was because all but one of ArcelorMittal's group companies in the Climate TRACE dataset had 'ArcelorMittal' in their original owner name. Meanwhile, only one of China Baowu Group's 14 group companies in the Climate TRACE dataset had 'China Baowu' in their name.

Overall, for the top 114 steel producers in 2021, Climate TRACE's asset-level production estimates include 81% of WorldSteel's reported production. For the top 20 steel producers, post-mapping desk research confirmed that only a marginal percent (1%) of missing production was due to detectable errors in mapping. However, it is always possible that errors in mapping exist, and simply were not detected during the mapping process, or in the validation checks conducted afterwards. In some cases, it was possible to positively confirm that mapping errors were not the cause of missing production. In the case of ArcelorMittal, missing production was due to the prioritization of high-emitting assets in Climate TRACE's dataset. Comparing numbers reported in ArcelorMittal's 2020 annual report, Climate TRACE captured 99.6% of ArcelorMittal's production derived from high-emitting blast furnaces, but only 48.7% of ArcelorMittal's production from lower-emitting electric arc furnaces.

In other cases, production was confirmed to be missing because the necessary inputs for estimating emissions for certain assets were not available for modeling. For example, Climate TRACE's production estimate for the 15th largest steel producer in 2021, Shougang Group, includes only 52.8% of Shougang's self-reported total. However, while both the GEM wiki and Shougang's website list these assets, the GEM wiki did not have plant capacities or any other information listed aside from ownership. The total production capacity for these plants. listed on Shougang's website, exceeds the total production missing from Climate TRACE's estimate. None of Shougang's mapped assets were incorrectly attributed during the mapping process. The final possible source of error is that production estimates were misattributed by Climate TRACE's production model. Results for Shougang Group are not consistent with that conclusion, but the scope of the current analysis cannot rule that out for the entire asset-level dataset. For a detailed discussion of how Climate TRACE estimates steel production, see the Climate TRACE steel methodology.

Similar validation checks were performed for the top 50 global cement producers, and top 50 oil and gas producers in Texas, with mapping results equivalent to those for steel. Altogether, these preliminary analyses suggest that Climate TRACE's mapping algorithm is potentially a robust, reliable method for generating time-efficient, up-to-date corporate network maps and company-level emissions estimates for the world's top emitting assets.

4. References

1) Clark, A. and Benoit, P. (2022) 'Greenhouse Gas Emissions from State-Owned Enterprises: A Preliminary Inventory'. Sipa Center on Global Energy Policy. Available at: https://www.energypolicy.columbia.edu/research/report/greenhouse-gas-emissions-state-own ed-enterprises-preliminary-inventory
2) Heede, R. (2014) 'Tracing anthropogenic carbon dioxide and methane emissions to fossil fuel and cement producers 1854–2010'. Climatic Change 122, 229–241. https://doi.org/10.1007/s10584-013-0986-y
3) Starr, D. (2016) "Just 90 Companies Are to Blame for Most Climate Change, This 'Carbon Accountant' Says." AAAS Articles DO Group. Available at: https://www.science.org/content/article/just-90-companies-are-blame-most-climate-change-carbon-accountant-says
4) GEM contributors (2022) 'Steel inventory data retrieved from: GEM Global Steel Plant Tracker'. Global Energy Monitor (GEM). Available at: https://globalenergymonitor.org/projects/global-steel-plant-tracker/.
5) McCarten, M., Bayaraa, M., Caldecott, B., Christiaen, C., Foster, P., Hickey, C., Kampmann, D., Layman, C., Rossi, C., Scott, K., Tang, K., Tkachenko, N., and Yoken, D. (2021) 'Global Database of Cement Production Assets'. Spatial Finance Initiative (SFI). Available at:Database of Cement Production Assets'. Spatial Finance Initiative (SFI). Available at: https://www.cgfi.ac.uk/spatial-finance-initiative/geoasset-project/geoasset-databases/ (Accessed: 26 August 2022).
6) Armstrong, T. *et al*. (2021) 'The Global Cement Report™'. 14. Tradeship Publications Ltd, UK. Available at: www.CemNet.com
7) Global Cement Directory (2022). Available at: https://www.globalcement.com/directory The Official Airline Guides Historical Flight Status Data (2022). Available at: https://www.oag.com/
8) Rystad Energy UCube (2021). Available with license at: https://www.rystadenergy.com/
9) GEM Contributors (2021) 'Coal reserves'. Global Energy Monitor. Available at: https://www.gem.wiki/w/index.php?title=Coal_reserves&oldid=239981
10) EIA: United States Energy Information Administration (2022). "Form EIA-860 detailed data with previous form data (EIA-860A/860B)" Available at:

https://www.eia.gov/electricity/data/eia860/
11) Lloyd's List Intelligence (2022). Available at: https://www.lloydslistintelligence.com/about-us/our-data
12) Light Metal Age (2022). Available at: https://www.lightmetalage.com/