Asset & Company-Level Ownership Methodology

Lee Gans[1], Brad Hall[2], Gavin McCormick[1], Ishan Saraswat[1], Gabriela Volpato[1], Krsna Raniga[1], Christy Lewis[1], Jeremy Freeman[1], Aaron Davitt[1], Lauren Schmeisser[3], Badr M'barek[4], Mason Phillpott[4], Matthew Jolleys[5], and Mark Powell[6]

1) WattTime, 2) Office of Al Gore, 4) RMI, 4) TransitionZero, 5) Hypervine Limited, 6) OceanMind.

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 CO_2 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 As a result, such reports are often out of date by the time they are published, and/or limited in the scope of their coverage.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 has generated the first ever independent, openly accessible asset and company-level emissions database – providing ownership information for 4,342 private companies and SOEs, located in 234 different countries and administrative regions. The most comprehensive coverage comes from the sectors that have been historically the most difficult to estimate – steel, oil, and gas. Climate TRACE has developed novel methods for generating up-to-date asset-level emissions estimates, as well as a mapping algorithm to match those assets – en masse – with large institutional ownership networks. The current discussion explains the methodology for the latter: Climate TRACE's corporate ownership mapping algorithm. In prior ownership studies, researchers have often had few methodological options other than downloading pre-aggregated, officially disclosed datasets, painstakingly researching companies and assets, one by one, or manually cleaning messy, crowdsourced data. While desk research and manually implemented quality controls were essential parts of the Climate TRACE mapping process, a substantially large proportion of it was done through automated, time-efficient methods that cross-reference information between available datasets. One of the obstacles in automated ownership mapping is the difficulty in testing its accuracy. For three sectors, Climate TRACE validated the results of the mapping algorithm by comparing company-level production estimates generated by Climate TRACE with external, self-reported industry and regulatory datasets. These initial results suggest that the mapping process generates similar estimates for production as compared to companies' self-reported datasets (adjusted for the proportion of the companies' assets that are covered by the asset-level emissions database). Climate TRACE's ownership data covers the 500 top emitting assets in 8 out of 23 Climate TRACE sectors, and at least the top 10 emitting assets in 3 sectors. In total, asset-level emissions estimates with ownership data are provided for 27.2% of Climate TRACE's global estimate for emissions in 2021.

2. Materials

Identification of company-level ownership occurred in two stages. First, datasets containing asset-level owners were aggregated into sector-specific datasets. For this analysis, an 'asset-level owner' is defined as the company whose name is attached to a single individual asset in a dataset. Once asset-level ownership datasets were aggregated, this data was input into a mapping algorithm to identify the asset-level owner's ultimate parent company, subsidiaries, joint ventures, and sibling companies. Assets that share an ultimate parent were then assigned to the same owner grouping to generate company-level emissions estimates. Emissions (and, where applicable, production) estimates for each owner grouping include the sum of all estimates for associated assets, in proportion to the owner group's percentage of ownership for each one.

2.1 Asset-level Ownership Datasets by Sector

For this section, the following asset-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

Steel. Assets 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 asset-level ownership data source for steel is the Global Energy Monitor (GEM) Wiki, an open database created by GEM and the Center for Media and Democracy that is updated yearly. GEM’s Global Steel Plant Tracker (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). The ‘owner name’ in the Climate TRACE dataset was derived from the ‘owner’ field on the GEM wiki where available, and the ‘parent’ field if not.

Cement. Assets were defined at the level of individual cement 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 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 1,538 of these plants. To ensure ownership data covered the top 500 emitting cement assets, desk research and industry reports were used to provide ownership information for an additional 79 plants (Armstrong *et al*., 2021; Global Cement Directory, 2021).

Electricity Generation. Assets 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 552 plants was derived from the GEM Wiki’s Global Coal Plant Tracker (GCPT) and Global Gas Plant Tracker (GGPT) (GEM, 2022). To ensure ownership coverage for the top 500 emitting assets, ownership information for 13 plants was derived through desk research. In some cases, there were different owners for specific units within an asset. 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. The 'owner name' in the Climate TRACE dataset was derived from the 'owner' field on the GEM wiki where available.

Oil and Gas Production and Transport. Assets were defined at the level of oil and gas fields, or macro-geological formations. Asset boundaries were derived from shapefiles included in an industry dataset licensed from Rystad Energy. Percent ownership for each asset was defined based on the owner's percent financial interest in the total oil and gas production for the asset. Rystad provides owner name and ownership percentage for most Climate TRACE assets. In some cases, ownership data was available only for subfields within assets. 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.

Oil and Gas Refining. Assets 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. Ownership for refineries was derived from desk research using company websites, government sources, and news articles.

International and Domestic Aviation. Assets were defined at the level of airports. Ownership was defined in terms of the percentage of the asset'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.

Coal Mining. Assets 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). The 'owner name' in the Climate TRACE dataset was derived from the 'parent' field on the GEM wiki where available, and the 'owner' field if not.

Solid Waste Disposal. Assets were defined at the level of individual landfills and dumpsites. 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 for solid waste assets were derived from desk research using company websites, government sources, and news articles.

Shipping. Assets were defined at the level of individual vessels. Ownership was defined in terms of percent financial interest in the asset as a piece of property. Ownership for shipping assets was derived from desk research using company websites, government sources, and news articles.

Enteric Fermentation. Assets were defined at the level of individual cattle feedlots. 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 for feedlots was derived from desk research using company websites, government sources, and news articles.

Table 1 summarizes the coverage of ownership information for the 11 sectors in which Climate TRACE provides asset-level emissions estimates and ownership data. The number of countries column shows how many countries contain at least one asset with an identified owner for the sector. For all sectors where more than 500 assets have identified owners, the data includes the top 500 emitting assets. For Solid Waste Disposal, the assets are among the top 500 and include the top 50. For Enteric Fermentation and Shipping, the assets are within the top 30. The "% Emissions" columns show the percentage of 2021 global emissions derived from assets with identified owners at three levels- "% Emissions Assets": emissions from all identified assets, "% Emissions Sector": emissions from the entire sector (asset-level + remaining country-level emissions), and "% Emissions Global": total sector emissions that Climate TRACE has ownership data for as a percentage of all Climate TRACE emissions.

Table 1 Asset and Emissions Coverage for Ownership Data by Sector

Sector	# Assets Ownership Identified	# Assets Emissions Estimated	# Assets w/ Ownership %	Years	# of Countries	% Emissions Asset	% Emissions Sector	% Emissions Global
Domestic Aviation	4618	4618	100.0%	2015-2022	174	100.0%	100.0%	0.5%
International Aviation	1982	1982	100.0%	2015-2022	233	100.0%	100.0%	0.4%
Oil and Gas Refining	591	592	99.8%	2021	103	99.5%	96.3%	1.8%
Oil and Gas Production and Transport	571	574	99.5%	2015-2021	91	99.6%	62.0%	5.8%
Coal Mining	2736	2754	99.3%	2021	65	99.5%	100.0%	2.3%
Steel	845	862	98.0%	2021	72	95.9%	79.7%	3.5%
Electricity Generation	565	601	94.0%	2021	41	97.2%	43.4%	10.0%
Cement	1617	2255	71.7%	2021	118	80.3%	60.6%	2.7%
Solid Waste Disposal	338	4498	7.5%	2021	29	42.6%	13.4%	0.2%
Enteric Fermentation	12	2388	0.5%	2021	1	8.6%	0.1%	0.003%
Shipping	25	45693	0.1%	2021	12	0.5%	0.3%	0.01%

The differences between sectors in terms of percentage of assets with identified ownership is due to differences in the availability of public and/or shareable licensed datasets, as well as time constraints. The GEM Wiki is a public dataset, while Rystad, SFI, and OAG data were licensed and published with permission for the specific purpose of this analysis. For oil and gas refining, solid waste disposal, enteric fermentation, and shipping, data was either not easily accessible or not permitted to be shared. However, because oil and gas refining constitutes such an outsize percentage of global emissions (1.8%), time was allocated toward the desk research necessary to manually aggregate a virtually complete dataset for 2021. In total, ownership data is provided for assets whose emissions account for 27.2% of total global emissions from all Climate TRACE sectors in 2021.

2.2 Company-Level Corporation Mapping Datasets

2.2.1 Automated Mass Mapping Datasets

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. Lists of all Wikidata entity types (called ‘instances’) and properties can be found here. 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. Although Wikidata is crowdsourced, datasets include citation links so that it is possible to track down the source for each data point.

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.

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 3.1 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.

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 3.1 below.

2.3 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. 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).

Oil and Gas Production and Transport. For 2021, 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 5.7% of global emissions for the sector overall and 9.2% of emissions from asset-level data 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 3.1.

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 2021 (Global Cement Directory, 2021).

3. Methods

3.1 Pre-Processing

Prior to mapping corporate networks, owner names from asset-level datasets, Wikidata, and GLEIF were consolidated, cleaned, and transformed into several standard formats for the purposes of matching. All pre-processing was completed using OpenRefine (see the OpenRefine User Guide for complete, detailed information).

Consolidation. Raw ownership data from all asset-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.

3.2 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.

3.3 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 for:

- Owners of the top 500 emitting assets
- Owner names from validation datasets that were among the top 200 global producers in the sector. If an entity is one of the top producers in the world, it is more likely that their assets are included in the asset-level dataset, even if the process did not return a corporate map or matched assets.

Desk research included custom, manual searches for the three 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.

3.4 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 and groupings 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. These criteria were applied to all sectors, with exceptions noted for steel and oil and gas production and transport. Companies with more than 50% ownership of an asset are listed as the sole "owner_grouping" for the asset in the Climate TRACE database. The other owners, their percent ownership, and their ultimate parents are listed in the "owner_name", "percentage_of_ownership", and "owner_direct_parent fields", respectively. Percent ownership refers to the company listed in the “owner_name field”. If more than one company is listed as a direct parent, they are the two largest parents or shareholders for the owner whose interest in the owner totals at least 50%. If one “owner_direct_parent” is listed, that parent’s ownership interest is greater than 50% (with sector-specific exceptions). For a single asset, the sum of all percent ownerships listed for owner_name’s that share the same owner_direct_parent constitutes the percentage ownership for that parent. Percentages for direct parents were calculated for each asset and used to determine the proper owner_grouping for the entire asset. In cases where no owner had a majority stake, the companies with the largest interests totaling at least 50% were listed as a single owner_grouping. If this string was more than 88 characters, companies with the largest interests totaling at least 33% were listed along with the word ‘Others’. For assets in which over 50% of the owners were individual persons, the owner is listed as ‘Unknown’ –even if their identity was listed in the original asset-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 marginal 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 hedge fund as an owner’s sole financier, the hedge fund 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 in countries outside the US, the owner_groupings for both oil and gas sectors list both the name of the SOE and the country in parenthesis (ex. Oil India (India)). The names in parentheses can be used to consolidate the dataset to look specifically at the global assets of sovereign nations, regardless of which specific SOE or agency is listed as the owner. Because it 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.

3.5 Depth of Analysis by Sector

Due to a combination of constraints related to data quality, availability, and time allocation, the ownership data for different sectors was processed at differing levels of analysis. The previous subsections describe all the steps that were conducted for the sectors that received the deepest level of analysis. However, Table 5 shows a breakdown of the depth of processing for each ownership sector. 'All steps' appears for sectors where every method described in sections 3.1-3.3 was implemented. 'Validation' appears for sectors where estimated actual production (or of percent ownership of production) from Climate TRACE was compared with an official industry or government dataset to check the accuracy of the matching, and coverage for each company's assets. The other terms in the table (ex. 'Consolidation') straightforwardly refers to the specific steps outlined sections 3.1-3.3.

Table 5 Depth of Analysis by Sector

Sector	Depth of Analysis
Domestic Aviation	Consolidation, OpenCorporates Reconciliation
International Aviation	Consolidation, OpenCorporates Reconciliation
Oil and Gas Refining	All steps
Oil and Gas Production and Transport	All steps, Validation
Coal Mining	Consolidation, Wikidata matching
Steel	All steps, Validation
Electricity Generation	All steps
Cement	Consolidation, Wikidata matching, Validation
Solid Waste Disposal	Consolidation (Most owners are townships)
Enteric Fermentation	Desk Research, Mapping
Shipping	Desk Research, Mapping

4. Selected Results

4.1 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).

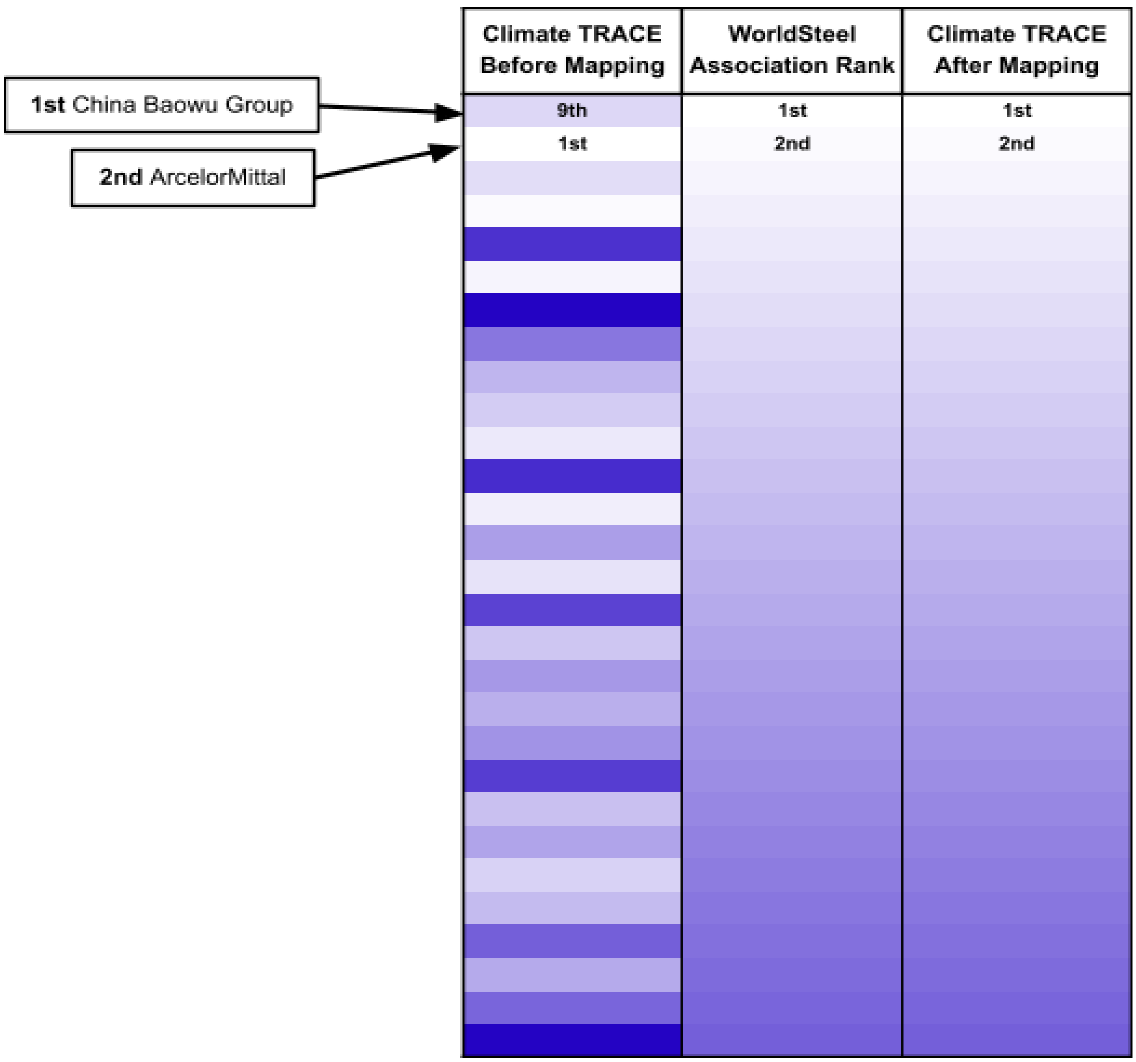

Figure 1 Ranking of 2021 Production Estimates from Climate TRACE Before and After Mapping. In the figure, China Baowu Group was ranked 1st for steel produced followed by ArcelorMittal based on self-reports to the WorldSteel Association (boxes with arrows on the left). However, prior to mapping each company's 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.

4.2 Oil and Gas Production and Transport

Recently, the New York Times published an article on a report from the [Private Equity Stakeholder Project]. The report details how private equity firms have been scaling up their acquisition of oil and gas assets in the United States (Tabuchi, 2021). The article describes the relative lack of transparency for private equity firms vs. publicly traded companies as a major reason for this trend. Due to public outcry over climate change, the highest emitting oil and gas assets are becoming perceived as risky investments for public companies. This creates an opportunity for private equity firms to buy these assets at low prices with relatively little public scrutiny. Meanwhile, publicly traded corporations can sell their high-emitting assets and exchange them with lower-emitting ones and count those offloaded assets as a reduction in their publicly reported emissions. Private equity firms often acquire these assets through becoming the seed investors behind small, independent companies that complete these purchases. As a result, assets backed by private equity firms receive less regulatory scrutiny than do those owned by traditional oil and gas conglomerates.

Climate TRACE's asset-level and ownership dataset for oil and gas production covers the years 2015-2021. During mapping, owners backed by private equity firms were identified. Figure 2 shows Climate TRACE's estimates for how the proportion of emissions generated by different types of oil and gas investors has changed since 2015, alongside Climate TRACE's estimates for how marginal emissions (emissions per unit of production) have changed since 2015. The black bar shows traditional oil and gas corporations, the gray bar shows small businesses (companies with <0.0001% interest in an asset's production) and individual persons or "smallholders", and the green bar shows private equity. The proportion of total USA oil and gas production emissions from private equity owned-assets has grown since 2015, while the proportion generated by smallholders has fallen. The proportion of emissions generated by traditional oil and gas corporations has remained relatively stagnant.

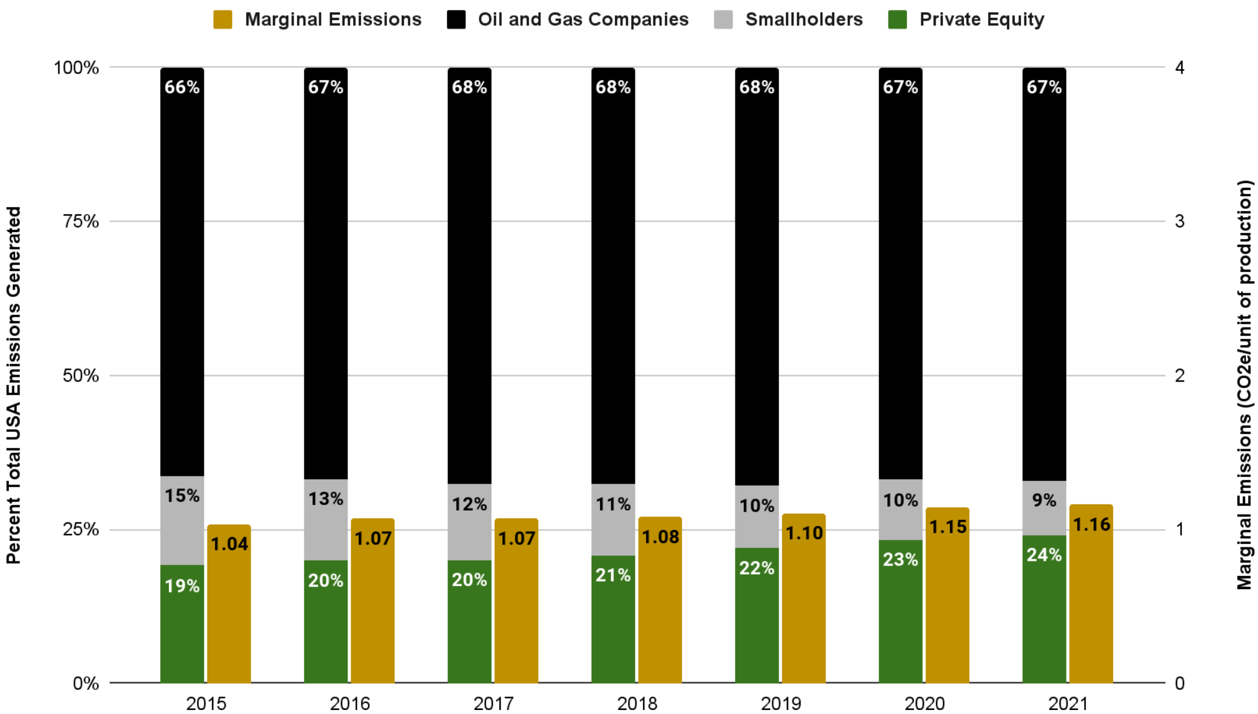

Figure 2 Proportion of Oil and Gas Production Emissions Generated by Private Equity vs. Smallholders vs. Traditional Oil and Gas Companies in the United States, 2015-2021. Left y-axis: Percent of Total USA Emissions Generated by ownership: Oil and Gas Companies (black bars), Smallholders (gray bars), and Private Equity (green bars). Right y-axis: Marginal Emissions (brown bars).

Although traditional oil and gas companies remain the largest emitters, these results are consistent with the concerns reported by the Private Equity Stakeholder Project and the New York Times. These data are consistent with the hypothesis that marginal emissions increase as higher-emitting assets pass from publicly traded companies to private equity investors. Between 2015 and 2021, the proportion of emissions generated by private equity assets increased 19%, while marginal emissions increased 12%.

5. Discussion, Limitations, and Future Directions

These preliminary analyses suggest Climate TRACE's mapping algorithm has great potential as a time-efficient, up-to-date, and reliable method for generating company-level emissions estimates. Because Climate TRACE's production estimates correspond well to self-reported production numbers, this creates a common reference point from which companies and SOE's who do not currently track their emissions can receive actionable information. Companies who currently report their emissions could also benefit by comparing their internal estimates to those of Climate TRACE. For investors, policymakers, and activists, the Climate TRACE company-level ownership dataset shows potential to provide insight into economic trends that drive changes in emissions. Results from the oil and gas sector may reflect a phenomenon colloquially referred to as ‘hot potato’, where high-emitting assets are passed around to avoid accountability for implementing real emission reductions. These results suggest when companies play ‘hot potato’ with their assets, they may not only end up reporting artificial reductions in their emissions, but their actions may actually result in emissions increasing. Future Climate TRACE analyses will explore this hypothesis further.

Although these results are promising, there are some important limitations. One is the lack of coverage for all assets and sectors. Another is that these analyses are preliminary and have yet to be replicated robustly – with just three sectors having been mapped at the deepest level of analysis. It should also be noted that, while many of the companies in the Climate TRACE dataset own cross-sector assets, mapping has not been completed between sectors, and unique parent entities may sometimes be identified by different name variants in different sectors. Finally, although these analyses provide unprecedented coverage of Chinese steel assets and international oil and gas production, the corporate mapping datasets still have limited coverage for companies and SOEs in the global south. Climate TRACE intends to increase coverage, replicate and validate more results, identify new mapping datasets, and perform cross-sector mapping in future analyses. Other future directions include: identifying how specific investment models and policies impact the generation of emissions by private companies and SOE’s.

6. 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-owned-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:

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
8. The Official Airline Guides Historical Flight Status Data (2022). Available at: https://www.oag.com/
9. Rystad Energy UCube (2021). Available with license at: https://www.rystadenergy.com/
10. GEM Contributors (2021) ‘Coal reserves’. Global Energy Monitor. Available at: https://www.gem.wiki/w/index.php?title=Coal_reserves&oldid=239981
11. Tabuchi, H. (2021) “Private Equity Funds, Sensing Profit in Tumult, Are Propping up Oil.” The New York Times, 13 Oct. 2021. Available at: https://www.nytimes.com/2021/10/13/climate/private-equity-funds-oil-gas-fossil-fuels.html