

Datenkatalog

Alles, was für das S&P-500-Nachhaltigkeitsranking gesammelt wurde –
einmal nach Quelle, einmal nach Datentyp

ETHack 2026

Stand aller Abrufe: 12. September 2026

Wozu dieser Katalog. Der Bestand ist an einem Tag aus 19 Quellen zusammengetragen worden. Damit er später wiederverwendbar ist, braucht es zwei Dinge: eine Liste, woher jeder Wert stammt und was sein Abruf kostet, und eine Liste, was die Werte eigentlich messen. Beide Tabellen sind *erzeugt*, nicht *geschrieben* – `scripts/06_katalog.py` liest sie aus den Daten. Kommt eine Quelle dazu, stimmt der Katalog beim nächsten Lauf von selbst.

Der Bestand in Zahlen

Quellen	19
Rohtabellen im Zwischenspeicher	24
Rohzeilen insgesamt	9 882 445
Umfang der Rohdaten	255,5 MB
Werte im gesammelten Datensatz	11 798
davon Firmen	503 von 503
davon Kennzahlen	45
abgedeckte Geschäftsjahre	2018–2026
Werte ohne Quellenangabe	0
Quellen mit Schlüsselpflicht	2

Inhaltsverzeichnis

Der Bestand in Zahlen	1
1 Aufbau des gesammelten Datensatzes	3
2 Sicht 1: nach Quelle	4
2.1 Rohdaten im Zwischenspeicher	6
3 Sicht 2: nach Datentyp	7
4 Belastbarkeit je Bereich	9
5 Was beim Weiterverwenden zu beachten ist	10
5.1 Die Kennzahlen sind nicht gleichwertig	10

5.2	Zwei Quellen mit Selbstselektion	10
5.3	Eine Lizenz, die Vorsicht verlangt	10
5.4	Geprüft, aber nicht aufgenommen	10
5.5	Gegenprobe mit den Team-Daten	11
5.6	Was der Katalog nicht enthält	11

6	Reproduktion	11
----------	---------------------	-----------

1 Aufbau des gesammelten Datensatzes

Der Datensatz liegt im Langformat vor: **eine Zeile je Firma, Jahr und Kennzahl**. Das ist absichtlich so und nicht als breite Tabelle, weil sich nur so die Herkunft an jedem einzelnen *Wert* führen lässt statt nur an der Spalte.

Der Unterschied ist nicht kosmetisch. Die CO₂-Zahl einer Firma kann am Schornstein gemessen sein, die der nächsten aus Brennstoffmengen gerechnet und die der dritten aus einem Geschäftsbericht abgeschrieben. Eine breite Tabelle würde diese drei Fälle in einer Spalte `co2` zusammenwerfen und die Unterscheidung verlieren – genau die Information, an der ein Ranking steht oder fällt.

Spalte	Inhalt
<code>ticker, company</code>	Firma
<code>gics_sector, gics_sub_industry</code>	Branchenzuordnung
<code>year</code>	Geschäftsjahr des Wertes
<code>metric, metric_label</code>	Kennzahl, technisch und lesbar
<code>value, unit</code>	Wert und Einheit
<code>direction</code>	−1 kleiner ist besser, +1 größer ist besser, 0 neutral
<code>axis</code>	A Kernnote, B Glaubwürdigkeit, S sozial, G Governance, sonst Bezugsgröße
<code>source_id, source_name</code>	Quelle
<code>source_url</code>	Abrufadresse
<code>source_access, source_license</code>	Zugangsbedingung und Lizenz
<code>retrieved_at</code>	Abrufdatum

Dateien. `data/out/dataset_long.parquet` und `.csv` enthalten alle Werte. `sources.json` und `metrics.json` sind die beiden Register. `companies.json` ist eine je Firma verdichtete Sicht, die die Web-Oberfläche direkt liest.

2 Sicht 1: nach Quelle

Je Quelle steht hier: ob ein Schlüssel nötig ist, unter welcher Lizenz sie steht, wie viele Rohzeilen im Zwischenspeicher liegen, für wie viele Firmen sie am Ende einen Wert liefert – und wo ihre Grenze liegt.

Quelle	Key	Lizenz	Rohzeilen	Firmen	Werte
EPA Greenhouse Gas Reporting Program https://data.epa.gov/efservice/PUB_DIM_FACILITY/ US-Anlagen ueber 25 000 t CO ₂ e, Berichtsjahre 2018-2023 Grenze: Reihe endet 2023; Direktemittenten von Lieferanten getrennt (sector_type)	nein	US-Regierungswerk, gemeinfrei	130 036	149	1 006
EPA Clean Air Markets Program Data https://api.epa.gov/easey/emissions-mgmt/emissions/apportioned/annual Kraftwerksblöcke der Handelsprogramme, 2019 bis laufendes Jahr unvollständig	ja	US-Regierungswerk, gemeinfrei	21 912	47	686
EPA eGRID https://www.epa.gov/egrid/download-data US-Kraftwerke, Berichtsjahr 2023 Grenze: nur aussagekräftig fuer Firmen, deren Geschäft Stromerzeugung ist	nein	US-Regierungswerk, gemeinfrei	12 239	45	90
SEC EDGAR XBRL frames https://data.sec.gov/api/xbrl/frames/ Geschäftsjahre 2016-2025 Grenze: fuer verschiedene Umsatz-Tags; Firmen taggen uneinheitlich	nein	oeffentlich	53 492	146	830
Science Based Targets initiative https://sciencebasedtargets.org/download/excel 15 605 Organisationen weltweit, Stand laufend fehlt fast ganz	nein	Nutzung mit Quellenangabe	15 605	241	757
EPA Toxics Release Inventory https://data.epa.gov/efservice/downloads/tri/mv_tri_basic_download/2023_US/csv/ US-Anlagen, Berichtsjahr 2023 Grenze: Gesamtmenge korreliert mit CO ₂ (rho 0,53); eigenständig ist der Krebserreger-Anteil	nein	US-Regierungswerk, gemeinfrei	36 108	194	582
OSHA Injury Tracking Application, Formular 300A https://www.osha.gov/itadata US-Betriebsstätten, 2023-2025 Grenze: DART misst auch Branchenstruktur; branchenrelativ vergleichen	nein	US-Regierungswerk, gemeinfrei	1 176 137	314	942
EPA ECHO Enforcement and Compliance History https://echo.epa.gov/files/echodownloads/echo_exporter.zip US-Anlagen, Konformitätshistorie drei Jahre mit GHGRP-Eintrag	nein	US-Regierungswerk, gemeinfrei	3 178 967	136	408
EIA Formular 923 (via PUDL und API) https://www.eia.gov/electricity/data/eia923/ US-Kraftwerke ab 1 MW, 2021-2025 Grenze: unabhngige Gegenprobe zu CEMS, nicht dieselbe Messkette	ja	US-Regierungswerk, gemeinfrei	264 315	–	–
Kommerzielle ESG-Risk-Ratings (Sustainalytics via Yahoo, gespiegelt) https://raw.githubusercontent.com/sburststein/ESG-Stock-Data/main/sp_esg_stock_data.csv 245 Firmen, Stand 2021 Grenze: eingefroren 2021; fliet in keine eigene Note ein	nein	unklar – nur als Vergleichsma	245	199	199
S&P-500-Konstituenten https://en.wikipedia.org/wiki/List_of_S%26P_500_companies 503 Mitglieder, heutiger Stand Grenze: nur der heutige Stand – fuer historische Jahre ueberlebensverzerrt	nein	CC BY-SA	503	–	–

Quelle	Key	Lizenz	Rohzeilen	Firmen	Werte
PUDL (Catalyst Cooperative) auf Zenodo https://s3.us-west-2.amazonaws.com/pudl.catalyst.coop/stable/ aufbereitete EIA-, EPA- und SEC-Daten, 1995 bis laufendes Jahr moeglich, dafuer entfaellt eigene Bereinigung	nein	Public Domain	1 162 085	–	–
SEC 10-K Anlage 21 (Konzernstruktur, via PUDL) https://s3.us-west-2.amazonaws.com/pudl.catalyst.coop/stable/core_sec10k__quarterly_exhibit_21_company_ownership_parquet 3,8 Mio. Tochtereintraege, 478 von 500 Indexmuettern Beteiligungsquote nur bei 99 117 Eintraegen	nein	Public Domain	3 800 706	–	–
SEC 10-K-Finanzkennzahlen (companyfacts, aufbereitet von Janis) https://data.sec.gov/api/xbrl/companyfacts/ Geschaeftsjahr 2024, 499 Indexfirmen berichtet	nein	oeffentlich	503	498	2 281
DOL Wage and Hour Division, Verfahren (aufbereitet von Janis) https://data.dol.gov/data-catalog/WHD/enforcement/WHD_enforcement.zip Feststellungsende 2022-2024 zu erscheinen. Gegenprobe mit eigenem exakten Abgleich: 127 statt 213 Firmen, Rangkorrelation der Fallzahlen 0,62 – die Team-Zuordnung zaehlt vermutlich Franchise-Betriebe unter dem Markennamen mit (McDonald's 142 gegen 64 Faelle)	nein	US-Regierungswerk, gemeinfrei	29 374	213	639
SEC Form SD, Konfliktmineralien (aufbereitet von Janis) https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&type=SD Meldungen 2022-2025	nein	oeffentlich	–	503	503
World Benchmarking Alliance, Unternehmensprofile 2026 (erhoben von Janis) https://www.worldbenchmarkingalliance.org/company-scoreboard 217 Indexfirmen, Bewertungsrunde 2026 ohne Streuung (216 von 217 = 0)	nein	CC BY 4.0 – Namensnennung: Wo	218	218	943
Wikidata, offizielle Websites (fuer Firmenlogos) https://query.wikidata.org/sparql Website ueber CIK (P5531 -> P856), Rest aus SEC Submissions und Marken der jeweiligen Firmen	nein	CC0	–	–	–
Eigene Berechnung dieser Pipeline scripts/02_analyse.py abgeleitet aus den oben genannten Quellen	nein	–	–	146	1 932
<i>Grenze: Herkunft der Eingangsgroessen steht jeweils beim Basiswert</i>					

Nur zwei Quellen verlangen eine Anmeldung, beide kostenlos und sofort. Der EPA-Schlüssel für die Clean Air Markets ist der wichtige: Ohne ihn greift DEMO_KEY mit zehn Anfragen je Stunde, mit ihm sind es tausend. Er ist auch der Schlüssel, der die zeitliche Lücke nach 2023 schließt. Der EIA-Schlüssel dient nur der unabhängigen Gegenprobe. Beide liegen in `.env`, die in `.gitignore` steht; eine Vorlage ohne Werte liegt als `.env.example` bei.

2.1 Rohdaten im Zwischenspeicher

Alle Abrufe werden als Parquet zwischengespeichert. Das macht die Auswertung reproduzierbar und schont die Server – zwei der Quellen brauchen mehrere Minuten je Abruf.

Tabelle	Zeilen	Spalten	MB	Jahre
sec10k_ex21	3 800 706	5	76.3	–
epa_echo	3 178 967	10	92.3	–
osha_ita	1 176 137	17	53.1	2023–2025
eia923_gen_fuel	727 126	9	12.4	2001–2025
eia_generation	181 340	4	2.5	2001–2025
out_eia__yearly_utilities	165 728	27	2.3	2001–2026
out_epacems__yearly_operational_characteristics	118 117	13	3.8	2000–2025
out_eia860__yearly_ownership	106 281	16	0.7	2001–2025
eia_api_facility_fuel	72 259	6	1.3	2019–2025
epa_facility	65 155	10	1.2	2018–2023
epa_emission	64 881	5	0.6	2018–2023
sec_revenue	53 492	5	0.8	2016–2025
core_epa__assn_eia_epacamd	44 833	7	0.2	2018–2024
epa_tri	36 108	122	4.8	2023–2023
dol_whd	29 374	14	1.6	–
sbt_i_targets	15 605	13	0.5	–
egrid_plant	12 239	11	0.6	2023–2023
campd_facility	11 051	8	0.1	2019–2026
campd_emission	10 861	7	0.2	2019–2026
epa_eia_crosswalk	10 716	3	0.0	2018–2024
team_harte_variablen	503	60	0.2	–
sp500_master	503	5	0.0	–
esg_snapshot	245	11	0.0	–
team_wba	218	11	0.0	–

3 Sicht 2: nach Datentyp

Dieselben Werte, gruppiert nach dem, was sie messen. Der Pfeil gibt die Richtung an: ↓ heißt, kleiner ist besser.

Kennzahl	technisch	Einheit	Richt.	Firmen	Jahre	Quelle
Emissionsmengen						
Absolute Treibhausgasmengen, wie sie gemeldet oder gemessen wurden.						
Scope-1-Emissionen	scope1_t	t CO2e	↓	149	2018–2023	EPA Greenhouse Gas Reporti
CO2 gemessen (CEMS)	campd_co2_t	t CO2e	↓	47	2019–2026	EPA Clean Air Markets Prog
Intensitaeten und Trends						
Emissionen bezogen auf eine Aktivitaetsgroesse, und ihre Entwicklung. Hier entscheidet die Wahl des Nenners.						
CO2-Intensitaet	co2_intensity	t CO2e je Mio. USD Umsatz	↓	146	2018–2023	Eigene Berechnung dieser P
CO2 je erzeugter MWh	t_co2_pro_mwh	t CO2e/MWh	↓	45	2023–2023	EPA eGRID
Trend der Intensitaet	intensity_cagr	%/Jahr	↓	127	2022–2023	Eigene Berechnung dieser P
Trend der absoluten Tonnen	absolute_cagr	%/Jahr	↓	130	2022–2023	Eigene Berechnung dieser P
Weitere Umweltwirkung						
Was neben CO2 in Luft, Wasser und Boden gelangt.						
Giftstofffreisetzung	tri_releases_lbs	lbs	↓	194	2023–2023	EPA Toxics Release Invento
davon krebserregend	tri_carcinogen_lbs	lbs	↓	194	2023–2023	EPA Toxics Release Invento
Arbeitssicherheit						
Die soziale Dimension als gemeldete Kennzahl statt als Selbstbeschreibung.						
Unfallrate DART	dart_rate	Faelle je 100 Vollzeitkraefte	↓	314	2025–2025	OSHA Injury Tracking Appli
Todesfaelle	osha_deaths	Anzahl	↓	314	2025–2025	OSHA Injury Tracking Appli
Regeltreue						
Dokumentiertes Verhalten gegenueber Umweltauflagen. Die einzige Achse, die nachweislich unabhaengig von der CO2-Intensitaet ist.						
Umweltstrafen	echo_penalties_usd	USD	↓	136	2025–2025	EPA ECHO Enforcement and C
Verstossquartale je Anlage	echo_nc_quarters_per_site	Quartale	↓	136	2025–2025	EPA ECHO Enforcement and C
Anlagen mit schwerem Verstoss	echo_significant	Anzahl	↓	136	2025–2025	EPA ECHO Enforcement and C
Ziele und Glaubwuerdigkeit						
Achse B des Modells. Wird getrennt ausgewiesen und nie in die Kernnote eingerechnet.						
SBTi-geprueftes Ziel	sbti_validated	ja/nein	↑	241	2026–2026	Science Based Targets init
Zwischenzieljahr	sbti_near_term_year	Jahr	–	209	2026–2026	Science Based Targets init
Netto-null-Jahr	sbti_net_zero_year	Jahr	–	66	2026–2026	Science Based Targets init

Kennzahl	technisch	Einheit	Richt.	Firmen	Jahre	Quelle
Zusage zurueckgezogen	sbti_commitment_removed	ja/nein	↓	241	2026–2026	Science Based Targets init
Intensitaet faellt, Tonnen steigen	intensity_illusion	ja/nein	↓	143	2022–2023	Eigene Berechnung dieser P
Basisjahr-Bequemlichkeit	base_year_ratio	Verhaeltnis	↓	130	2022–2023	Eigene Berechnung dieser P
Arbeitsrecht						
Behoerdlich festgestellte Lohnverstoesse. Firmen ohne zuordenbaren Fall fehlen, statt als Null zu erscheinen.						
Lohnverfahren 2022-2024	whd_cases	Anzahl	↓	213	2024–2024	DOL Wage and Hour Division
nachgezahlte Loehne	whd_backwages_usd	USD	↓	213	2024–2024	DOL Wage and Hour Division
betroffene Beschaeftigte	whd_employees	Anzahl	↓	213	2024–2024	DOL Wage and Hour Division
Externe Bewertungen, offen lizenziert						
WBA-Benchmarks unter CC BY 4.0. Anders als der kommerzielle Snapshot ist die Methodik einsehbar – bewertet wird trotzdem Offenlegung, nicht Wirkung.						
WBA Qualitaet des Transitionsplans	wba_tpq	0-5	↑	218	2026–2026	World Benchmarking Allianc
WBA Beitrag zur Transition	wba_ctt	0-2	↑	218	2026–2026	World Benchmarking Allianc
WBA Social Benchmark	wba_social	0-100	↑	218	2026–2026	World Benchmarking Allianc
WBA Nature Benchmark	wba_nature	0-100	↑	71	2026–2026	World Benchmarking Allianc
WBA Just Transition	wba_just_transition	0-100	↑	218	2026–2026	World Benchmarking Allianc
Finanzkennzahlen						
Aus den 10-K-Berichten, als Bezugsgroessen fuer Intensitaeten und fuer die Portfoliofrage – keine Nachhaltigkeitsbewertung.						
Nettogewinn	net_income_usd	USD	–	446	2024–2024	SEC 10-K-Finanzkennzahlen
Bilanzsumme	total_assets_usd	USD	–	498	2024–2024	SEC 10-K-Finanzkennzahlen
Finanzschulden	total_debt_usd	USD	–	391	2024–2024	SEC 10-K-Finanzkennzahlen
operativer Cashflow	operating_cf_usd	USD	–	482	2024–2024	SEC 10-K-Finanzkennzahlen
Sachinvestitionen	capex_usd	USD	–	330	2024–2024	SEC 10-K-Finanzkennzahlen
Forschung und Entwicklung	rnd_usd	USD	–	134	2024–2024	SEC 10-K-Finanzkennzahlen
Lieferkette						
Ob eine Firma Konfliktmineralien meldet. Die Meldepflicht allein ist noch kein Befund ueber die Herkunft.						
meldet Konfliktmineralien (Form SD)	sd_conflict_minerals_filerja/nein		–	503	2025–2025	SEC Form SD, Konfliktminer
Ergebnis des Modells						
Kein Einzelplatz, sondern die Spanne ueber alle Methodenkombinationen.						
Rangband untere Grenze	rank_p10	Perzentil	↑	143	2023–2023	Eigene Berechnung dieser P

Kennzahl	technisch	Einheit	Richt.	Firmen	Jahre	Quelle
Rangband Median	rank_p50	Perzentil	↑	143	2023–2023	Eigene Berechnung dieser P
Rangband obere Grenze	rank_p90	Perzentil	↑	143	2023–2023	Eigene Berechnung dieser P
Vergleichsmassstab						
Fremdbewertung, ausschliesslich zum Gegenhalten. Fliesst in keine eigene Note ein.						
kommerzielles ESG-Risiko	esg_risk_total	Punkte	↓	199	2021–2021	Kommerzielle ESG-Risk-Rati
Bezugs- und Guetegroessen						
Nenner, Zaehlwerte und Konfidenzangaben, die keine Bewertung sind.						
Umsatz	revenue_musd	Mio. USD	–	146	2018–2023	SEC EDGAR XBRL frames
zugeordnete Anlagen	n_facilities	Anzahl	–	143	2022–2023	EPA Greenhouse Gas Reporti
Zuordnungskonfidenz	match_confidence	0-1	↑	143	2022–2023	Eigene Berechnung dieser P
Kraftwerke mit Messung	campd_plants	Anzahl	–	47	2019–2026	EPA Clean Air Markets Prog
eGRID-Kraftwerke	egrid_plants	Anzahl	–	45	2023–2023	EPA eGRID
TRI-Anlagen	tri_facilities	Anzahl	–	194	2023–2023	EPA Toxics Release Invento
gemeldete Betriebsstaetten	osha_sites	Anzahl	–	314	2025–2025	OSHA Injury Tracking Appli

4 Belastbarkeit je Bereich

Nicht jede Zahl trägt ein Urteil. Je Firma und Bereich wird deshalb geprüft, ob das Material stark genug ist. Kennzahlen, die denselben Sachverhalt beschreiben, bilden eine *Familie* und zählen einmal, mit dem Gewicht ihrer stärksten vorhandenen Kennzahl. Das Gewicht drückt Beweiskraft aus, nicht inhaltliche Wichtigkeit: 1,0 gemessen oder behördlich festgestellt, 0,8 gesetzlich gemeldet, 0,5 bis 0,7 abgeleitet oder unsicher zugeordnet, 0,3 bis 0,4 Bewertung von Offenlegung, 0,2 bloßer Meldestatus.

Belastbar ist ein Bereich bei mindestens zwei unabhängigen Familien, davon einer mit Gewicht ab 0,8, und einer Summe ab 1,5. **Eingeschränkt** bei einem Anker oder einer Summe ab 1,0, sonst **dünn**. Ein gemeinsames Ranking verlangt Belastbarkeit in allen drei Bereichen und mindestens sechs solche Firmen je Branche. Regeln und Werte stehen in `pipeline/reliability.py`, das Ergebnis je Firma in `belastbarkeit.csv`.

Familie	Gewicht	Firmen	Beweisart
E – Umwelt			
Treibhausgase	1,0	149	am Schornstein gemessen (CEMS); behördlich gemeldet (GHGRP); gemeldet, auf Umsatz bezogen
Erzeugungseffizienz	0,8	45	gemessen und gemeldet, nur Stromerzeuger
Schadstofffreisetzung	0,8	194	behördlich gemeldet (TRI)
Emissionstrend	0,5	130	abgeleitet aus mindestens drei Jahren
Naturbewertung	0,4	71	Bewertung veröffentlichter Angaben (WBA)

Familie	Gewicht	Firmen	Beweisart
Klimaziele	0,3	345	Bewertung des Transitionsbeitrags (WBA); Bewertung des Transitionsplans (WBA); Selbstauskunft, extern geprüft (SBTi)
S – Soziales			
Arbeitssicherheit	1,0	314	gesetzliche Meldung je Betrieb (OSHA)
Lohnrecht	0,7	213	behördlich festgestellt, Namenszuordnung unsicher
Sozialbewertung	0,4	218	Bewertung veröffentlichter Angaben (WBA)
G – Unternehmensführung			
Umweltregeltreue	1,0	136	behördlich festgestellt (ECHO)
Umgang mit Zusagen	0,5	241	protokollierter Rückzug (SBTi)
Greenwashing-Signale	0,4	143	abgeleitet aus Emissionsreihe
Lieferkettenpflicht	0,2	503	nur Meldestatus (Form SD)

5 Was beim Weiterverwenden zu beachten ist

5.1 Die Kennzahlen sind nicht gleichwertig

Drei Arten von Werten stehen nebeneinander und dürfen nicht vermischt werden:

- **Gemessen.** CAMPD misst am Schornstein, ECHO und OSHA sind behördlich festgestellt beziehungsweise gesetzlich vorgeschrieben.
- **Gemeldet.** GHGRP, TRI und SBTi beruhen auf Selbstauskunft – teils geprüft, teils nicht.
- **Berechnet.** Alles mit der Quelle „Eigene Berechnung“ ist abgeleitet. Die Herkunft der Eingangsgrößen steht jeweils beim Basiswert.

5.2 Zwei Quellen mit Selbstselektion

SBTi und der kommerzielle ESG-Snapshot enthalten nur Firmen, die sich gemeldet haben beziehungsweise die der Anbieter bewertet. Dass die Energiebranche bei SBTi zu null Prozent vertreten ist, ist selbst eine Information – aber keine, die man als Bewertung verrechnen darf.

5.3 Eine Lizenz, die Vorsicht verlangt

Der ESG-Snapshot ist über ein öffentliches Spiegel-Repository bezogen; die ursprüngliche Lizenzlage ist unklar. Er wird deshalb ausschließlich als Vergleichsmaßstab verwendet und fließt in keine eigene Note ein. Für eine Weitergabe des Datensatzes sollte er ausgenommen werden.

5.4 Geprüft, aber nicht aufgenommen

Mycelium. Im Team liegt eine Zuordnung von 485 Indexfirmen zu Mycelium-Emissionsprofilen, mit getrennt ausgewiesenem geschätzten Anteil – inhaltlich wertvoll, gerade für Firmen ohne Anlagendaten. Die Nutzungsbedingungen (mycelium.global/emissions-data-terms-of-use) untersagen jedoch, einen wesentlichen Teil der Datenbank „as a dataset, feed, or downloadable archive“ weiterzugeben, und zwar ausdrücklich auch für nicht-kommerzielle Zwecke. Systematisches Abrufen verlangt eine Lizenz. Die Werte fließen deshalb nicht in diesen Datensatz und nicht in die Downloads. Für eine Nutzung wäre die akademische API-Lizenz der richtige Weg.

Doppelte Behördenwerte. Die Team-Datei `sp500_harte_variablen.csv` enthält auch eigene Aufbereitungen von GHGRP, TRI und OSHA. Diese Quellen stehen bereits aus eigenem Abruf im Datensatz; übernommen wurden aus der Team-Datei nur die Größen, die sonst fehlen: Finanzkennzahlen, Lohnverfahren und Konfliktmineral-Meldungen. Die doppelten Werte eignen sich als Gegenprobe der Zuordnung.

5.5 Gegenprobe mit den Team-Daten

Janis hat die EPA-Treibhausgasdaten unabhängig aufbereitet, über die Konzernzuordnung der EPA selbst. Der Vergleich mit der eigenen Anlagenzuordnung deckte vier Fehler in der eigenen Tochtertabelle auf: National Grid und Avangrid standen bei Eversource (Faktor 25 zu hoch), Idaho Power bei Ameren, Tampa Electric bei Sempra. Außerdem wertete der unscharfe Namensabgleich Teilmengen als Volltreffer – aus „US Steel Corp“ wurde nach dem Streichen von Füllwörtern „steel“ und damit Steel Dynamics.

Nach der Korrektur stimmen beide Aufbereitungen für 122 gemeinsame Firmen mit einer Rangkorrelation von 0,987 überein, das mittlere Verhältnis liegt bei genau 1,00. Die größte verbleibende Abweichung ist PG&E: Die eigene Zuordnung rechnet die Tochter Pacific Gas & Electric mit ein, die Team-Datei nicht.

Bei den Lohnverfahren ist die Übereinstimmung schwächer: Ein eigener exakter Namensabgleich findet 127 statt 213 Firmen, die Fallzahlen korrelieren mit 0,62. Die Team-Zuordnung zählt vermutlich Franchise-Betriebe unter dem Markennamen mit. Der Datensatz führt die Team-Werte, der Quellenhinweis nennt die Abweichung.

5.6 Was der Katalog nicht enthält

67 der 503 Indexmitglieder haben zu keiner Kennzahl einen Wert, überwiegend Finanzdienstleister. Sie stehen in `companies.json` unter `withoutData` – ausdrücklich als „nicht bewertbar“ und nicht als Null.

Ebenfalls nicht enthalten: Scope 2 und Scope 3. Dafür existiert bis heute keine freie Quelle mit brauchbarer Abdeckung. Die kalifornische Offenlegungspflicht ändert das ab November 2026 für Scope 1 und 2 und ab 2027 für Scope 3.

6 Reproduktion

<code>cp .env.example .env</code>	Schlüssel eintragen (zwei Stück, beide kostenlos)
<code>scripts/01_fetch.py</code>	alle Quellen ziehen, als Parquet cachen
<code>scripts/02_analyse.py</code>	Zuordnung, Monte Carlo, beide Perioden
<code>scripts/05_neue_dimensionen.py</code>	Abdeckung und Unabhängigkeit der Zusatzachsen
<code>scripts/07_dataset.py</code>	gesammelten Datensatz und Web-Daten schreiben
<code>scripts/06_katalog.py</code>	diesen Katalog erzeugen

Einzelne Quelle neu ziehen: `scripts/01_fetch.py <name> --force`.