EPA Greenhouse Gas Search API For Automation

# Overview

The EPA provides an API for automated access to the data stored in the EnviroFacts database. An automated process may utilize the API to retrieve the data required to perform reporting and analysis, or simply to store the data in another location. The latter is the objective of our project. This document provides step-by-step instructions for downloading all of the data specifically related to the Greenhouse Gas Reporting Program (GHGRP) using the API.

The EPA actually provides two different API’s for access to its data. One is the EnviroFacts Data Service API, which crams all of the information needed to build a query into the URL. It is documented at the following locations:

<https://www.epa.gov/enviro/envirofacts-data-service-api>

<https://www.epa.gov/enviro/envirofacts-data-service-api-v1>

<https://www.epa.gov/enviro/web-services>

The other is the DMAP GraphQL API, which is based on the GraphQL standard, and is documented here:

<https://www.epa.gov/enviro/dmap-graphql-api>

Note: The tables in the database are categorized by subject area.

# Procedure

## Step 1: Get Table List

Fetch a list of all the tables in the database that are related to GHGRP, along with the metadata from the subject area associated with each of them. Submit an HTTP GET request to the following URL:

<https://data.epa.gov/dmapservice/metadata.qb_subject_area_tables/active/equals/true/join/metadata.qb_subject_areas/subject_area_id/equals/subject_area_id/program_code/equals/ghg/sort/title/base>

The resulting data will appear as follows:

{

"data": {

"metadata\_\_qb\_subject\_area\_tables": [

{

"active": **true**,

"subject\_area\_id": 23,

"table\_id": 251,

"table\_name": "w\_fuels\_for\_large\_ext\_comb",

"table\_schema": "ghg",

"title": "A collection of Data Elements Reported for Large External Combustion Units",

"metadata.qb\_subject\_areas.active": **true**,

"additional\_info": "<h5 style=\"color: #981b1e\">All data tables for 2011-2014 present emissions <u>BOTH</u> in mt (as calculated by EPA from reported data) <u>AND</u> in units of mt of Carbon Dioxide Equivalent (CO<sub>2</sub>e).<br><br><u>For 2011-2012, emissions were reported in units of mt CO<sub>2</sub>e using GWP's from the IPCC's Second Assessment Report.</u><br><br><u>For 2013-2014, emissions were reported in units of mt CO<sub>2</sub>e using GWP's from the IPCC's Fourth Assessment Report.</u><br><br>Emissions presented in the table W\_SUBPART\_LEVEL\_INFORMATION span the full time series (2011 on) and are in units of mt for all years.</h5><p>Data tables for Petroleum and Natural Gas Systems are generally provided separately for Reporting Year (RY) 2011-2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, and 2023. The reporting requirements for Petroleum and Natural Gas Systems differ between these groups of years. The table named EF\_W\_EMISSIONS\_SOURCE\_GHG contains each facility's total reported emissions for each industry segment and reporting source reported under Subpart W across the entire time-series (RY11 onwards).</p>",

"description": "Identifying a collection of data elements containing the total annual emissions of each greenhouse gas (GHG) listed in Table A-1 of 40 CFR 98 Mandatory Reporting of Greenhouse Gases reported under this Subpart, expressed in metric tons.)",

"group\_id": 7,

"program\_code": "ghg",

"subject\_area": "Subpart W - Petroleum and Natural Gas Systems (RY 2011-2014)",

"metadata.qb\_subject\_areas.subject\_area\_id": 23,

"metadata.qb\_subject\_areas.table\_schema": "ghg"

},

{

"active": **true**,

"subject\_area\_id": 23,

"table\_id": 303,

"table\_name": "w\_fuels\_for\_large\_int\_comb",

"table\_schema": "ghg",

"title": "A collection of Data Elements Reported for Large Internal Combustion Units",

"metadata.qb\_subject\_areas.active": **true**,

"additional\_info": "<h5 style=\"color: #981b1e\">All data tables for 2011-2014 present emissions <u>BOTH</u> in mt (as calculated by EPA from reported data) <u>AND</u> in units of mt of Carbon Dioxide Equivalent (CO<sub>2</sub>e).<br><br><u>For 2011-2012, emissions were reported in units of mt CO<sub>2</sub>e using GWP's from the IPCC's Second Assessment Report.</u><br><br><u>For 2013-2014, emissions were reported in units of mt CO<sub>2</sub>e using GWP's from the IPCC's Fourth Assessment Report.</u><br><br>Emissions presented in the table W\_SUBPART\_LEVEL\_INFORMATION span the full time series (2011 on) and are in units of mt for all years.</h5><p>Data tables for Petroleum and Natural Gas Systems are generally provided separately for Reporting Year (RY) 2011-2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, and 2023. The reporting requirements for Petroleum and Natural Gas Systems differ between these groups of years. The table named EF\_W\_EMISSIONS\_SOURCE\_GHG contains each facility's total reported emissions for each industry segment and reporting source reported under Subpart W across the entire time-series (RY11 onwards).</p>",

"description": "Identifying a collection of data elements containing the total annual emissions of each greenhouse gas (GHG) listed in Table A-1 of 40 CFR 98 Mandatory Reporting of Greenhouse Gases reported under this Subpart, expressed in metric tons.)",

"group\_id": 7,

"program\_code": "ghg",

"subject\_area": "Subpart W - Petroleum and Natural Gas Systems (RY 2011-2014)",

"metadata.qb\_subject\_areas.subject\_area\_id": 23,

"metadata.qb\_subject\_areas.table\_schema": "ghg"

},

…

}

If information only about the subject areas is desired, submit an HTTP GET request to this URL:

<https://data.epa.gov/dmapservice/metadata.qb_subject_areas/program_code/equals/ghg/active/equals/true/sort/subject_area/base>

The resulting data appears as follows:

{

"data": {

"metadata\_\_qb\_subject\_areas": [

{

"active": **true**,

"additional\_info": **null**,

"description": "Address and other location information about the facility.",

"group\_id": 6,

"program\_code": "ghg",

"subject\_area": "Facility Information",

"subject\_area\_id": 6,

"table\_schema": "ghg"

},

{

"active": **true**,

"additional\_info": **null**,

"description": "Contains information on greenhouse gas emissions, detailing the specific types of gas reported by facilities.",

"group\_id": 6,

"program\_code": "ghg",

"subject\_area": "Gas Information",

"subject\_area\_id": 8,

"table\_schema": "ghg"

},

…

}

If information about subject groups is desired, submit an HTTP GET request to this URL:

<https://data.epa.gov/dmapservice/metadata.qb_subject_area_groups/program_code/equals/ghg/is_ez/equals/false/sort/group_name/base>

The resulting data appears as follows:

{

"data": {

"metadata\_\_qb\_subject\_area\_groups": [

{

"display\_name": **true**,

"group\_id": 7,

"group\_name": "Detailed Subpart Subjects",

"is\_ez": **false**,

"program\_code": "ghg",

"sort\_order": 2,

"table\_schema": "ghg"

},

{

"display\_name": **true**,

"group\_id": 6,

"group\_name": "Summary Subjects",

"is\_ez": **false**,

"program\_code": "ghg",

"sort\_order": 1,

"table\_schema": "ghg"

}

]

},

"status": "COMPLETED"

}

## Step 2: Record Count

For each of the tables in the list, determine the number of records in the table. Submit an HTTP POST request to the following URL:

<https://data.epa.gov/dmapservice/query>

With the following body:

**query** aggregateCount {

ghg\_\_rlps\_ghg\_emitter\_facilities {

aggregate {

count

}

}

}

Replace the table name, rlps\_ghg\_emitter\_facilities, with the name of each target table, in turn. When submitting a request from Postman, select the option for GraphQL. For this particular table, the output will appear as follows:

{

"data": {

"ghg\_\_rlps\_ghg\_emitter\_facilities": [

{

"count": 107329

}

]

},

"status": "COMPLETED"

}

If the size of the data set exceeds the limit, or if the amount of time required to download the entire data set exceeds the limit, then pagination may be necessary. See step 4, below.

## Step 3: Get Table Contents

For each of the tables in the list, fetch the data. Submit an HTTP POST request to the following URL:

<https://data.epa.gov/dmapservice/query>

With the following body:

**query** fieldsQuery {

ghg\_\_rlps\_ghg\_emitter\_facilities {

\_\_all\_columns\_\_

}

}

Replace the table name, rlps\_ghg\_emitter\_facilities, with the name of each target table, in turn. When submitting a request from Postman, select the option for GraphQL. For this particular table, the output will appear as follows:

{

"data": {

"ghg\_\_rlps\_ghg\_emitter\_facilities": [

{

"facility\_id": 1000001,

"facility\_name": "PSE Ferndale Generating Station",

"address1": "5105 LAKE TERRELL ROAD",

"address2": **null**,

"city": "FERNDALE",

"state": "WA",

"state\_name": "WASHINGTON",

"zip": "98248",

"county\_fips": "53073",

"county": "WHATCOM COUNTY",

"latitude": 48.828707,

"longitude": -122.685533,

"year": 2010,

"cems\_used": **null**,

"parent\_company": "EMPECO IV, LLC AND USPF II FERNDALE HOLDINGS, LLC (74.33298%); DIAMOND GENERATING CORPORATION (14.00002%); TENASKA ENERGY, INC. AND TENASKA ENERGY HOLDINGS, LLC (11.667%)",

"add\_naics\_code": **null**,

"cogen\_unit\_emm\_ind": "Y",

"plant\_code": **null**,

"primary\_naics": "221112",

"secondary\_naics": **null**,

"epa\_verified": **null**

},

{

"facility\_id": 1000001,

"facility\_name": "PSE Ferndale Generating Station",

"address1": "5105 LAKE TERRELL ROAD",

"address2": **null**,

"city": "FERNDALE",

"state": "WA",

"state\_name": "WASHINGTON",

"zip": "98248",

"county\_fips": "53073",

"county": "WHATCOM COUNTY",

"latitude": 48.828707,

"longitude": -122.685533,

"year": 2011,

"cems\_used": **null**,

"parent\_company": "EMPECO IV, LLC AND USPF II FERNDALE HOLDINGS, LLC (74.33298%); DIAMOND GENERATING CORPORATION (14.00002%); TENASKA ENERGY, INC. AND TENASKA ENERGY HOLDINGS, LLC (11.667%)",

"add\_naics\_code": **null**,

"cogen\_unit\_emm\_ind": "Y",

"plant\_code": **null**,

"primary\_naics": "221112",

"secondary\_naics": **null**,

"epa\_verified": **null**

},

…

}

If you wish to receive the data in CSV format, instead of JSON, modify the URL as follows:

<https://data.epa.gov/dmapservice/query/csv>

## Step 4: Pagination

The API has a current limit of returning a maximum of 100,000,000 records. Furthermore, there is a hard cutoff of 15 minutes for a query. Any query exceeding that limit will not return results. If it isn’t practical to download all of the data with one request, then pagination may be necessary.

To paginate through results, the limit and offset parameters can be supplied.

query paging {

ghg\_\_rlps\_ghg\_emitter\_facilities (limit: 100, offset: 200) {

\_\_all\_columns\_\_

}

}

In this query, a maximum of 100 records will be returned. The first 200 matching records in the rlps\_ghg\_emitter\_facilities table will be skipped and the next 100 records will be returned.

## Step 5: Cross Program Facility Data

There is one more table we will need that does not appear in the list of subject area tables for GHGRP because it does not belong to any particular subject area. That is pub\_dim\_facility. This table contains facility data for all of the programs that store data in EnviroFacts, and stands above all of the other tables in the hierarchy established by the data model, as shown here:

<https://enviro.epa.gov/envirofacts/metadata/model/ghg?parentNode=4246>

The table structure is documented here:

<https://enviro.epa.gov/envirofacts/metadata/table/ghg/pub_dim_facility>

<https://enviro.epa.gov/enviro/GHG_NAV_TOOL_V2.get_step3?table_name=PUB_DIM_FACILITY>

There are two different ways to download its contents. One is to use the EnviroFacts Data Service API as follows:

<https://data.epa.gov/dmapservice/ghg.pub_dim_facility/csv>

The other is to use the DMAP GraphQL API as follows:

<https://data.epa.gov/dmapservice/query/csv>

With the following body:

**query** fieldsQuery {

ghg\_\_pub\_dim\_facility {

\_\_all\_columns\_\_

}

}