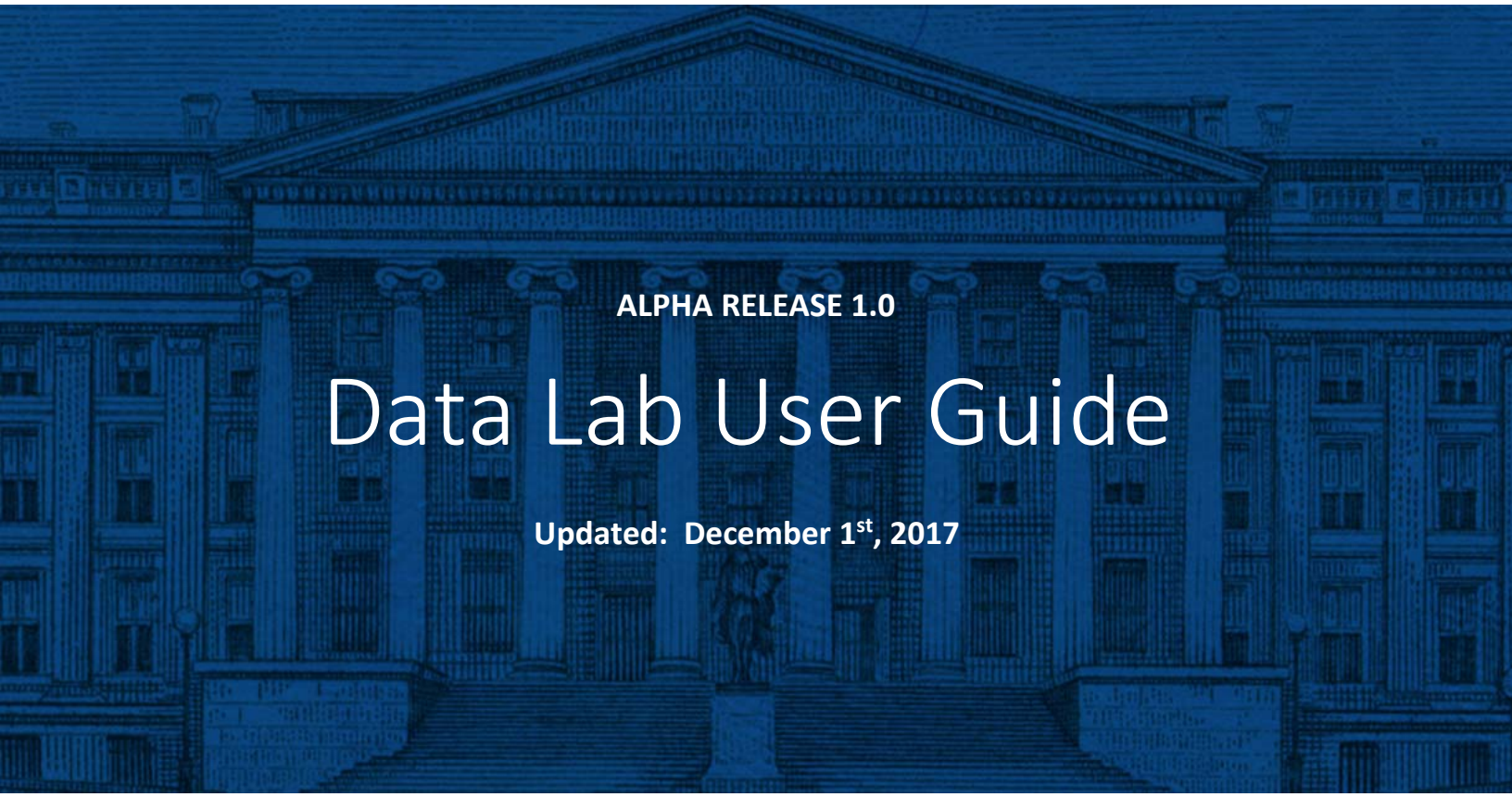


ALPHA RELEASE 1.0

Data Lab User Guide

Updated: December 1st, 2017

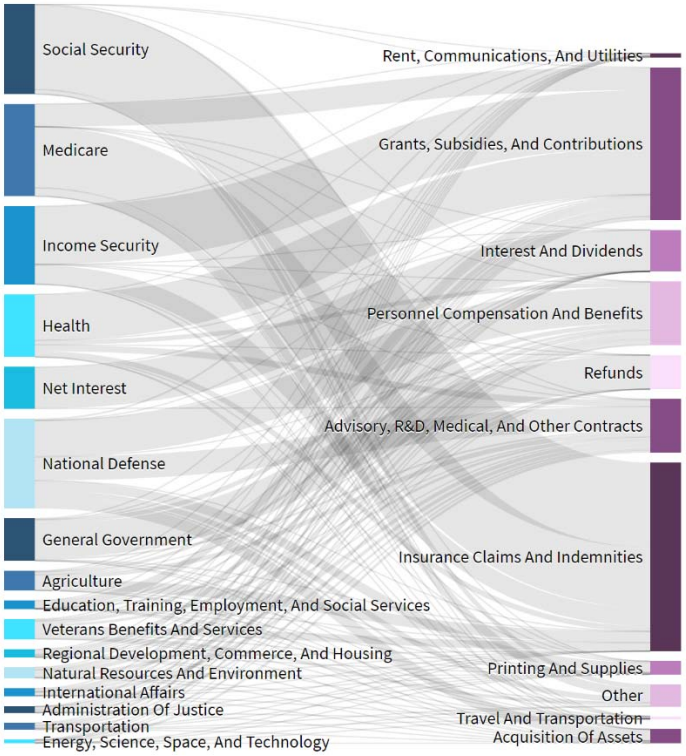
USASPENDING.gov
DATA LAB

Contents

Data Tools Developer Notes

The tools and visualizations that were published as part of Data Lab Alpha Release were built by our team using the following methodology.

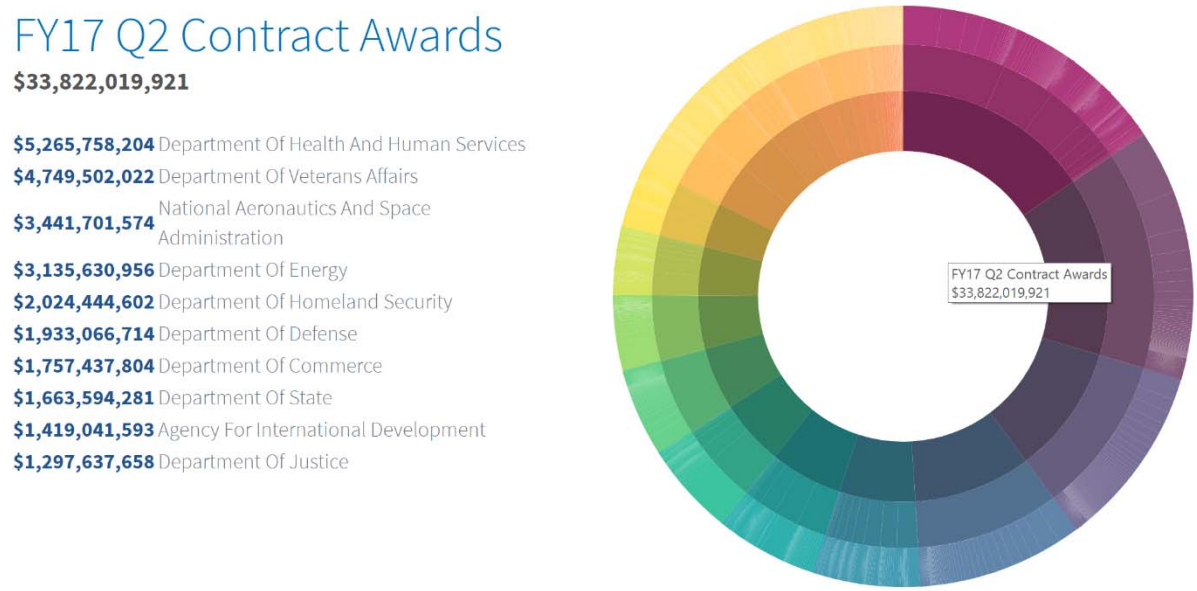
1. **BUDGET OBJECT CLASS SANKEY**



- Endpoint used: **TAS Categories**
- We grouped and aggregated using the following fields:
    - obligations_incurred_by_program_object_class_cpe, object_class.major_object_class_name, and
    - treasury_account.budget_function_title
- The twenty official congressional budget functions were rolled up into sixteen budget function categories. We combined budget functions which had similar infrastructure needs and/or similar missions.
- The five major object classes were broken down into eleven more specific classes, each of which focuses on highlighting specific budget item categories. These categories were created to be more descriptive and meaningful than the original five major object class categories, and less overwhelming than the thirty-five official object class categories.

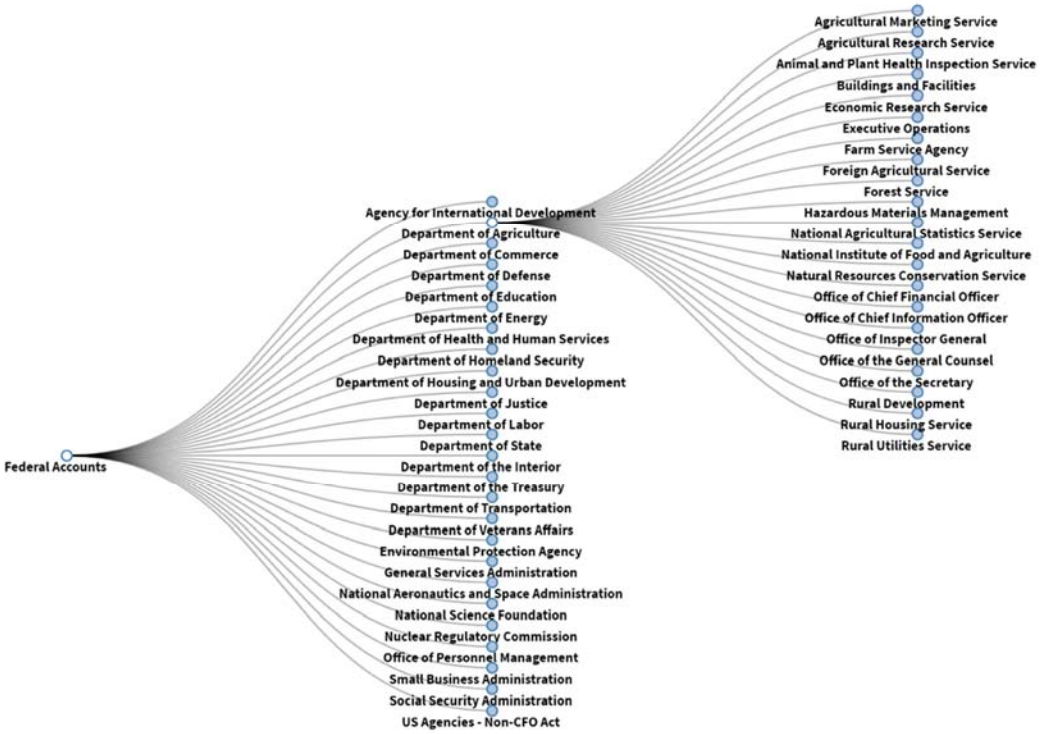
USASPENDING.gov
DATA LAB

2. CONTRACT EXPLORER SUNBURST



- Endpoints used: **Awards, Transactions**
- These endpoints were merged on piid. The piid field is available using a POST request on the Transactions endpoint.
- To create the visualization, the Awards endpoint data was subset by selecting records that correspond to contracts, designated by being of type: purchase order, delivery order, or definitive contract. From these records, we grouped and aggregated on:
    - funding_agency.toptier_agency.name
    - funding_agency.subtier_agency.name
    - recipient.recipient_name, and
    - total_obligation
- To create the information panel, the merged data set was grouped and aggregated on:
    - funding_agency.toptier_agency.name
    - funding_agency.subtier_agency.name
    - recipient.recipient_name
    - contract_data.product_or_service_code, and
    - total_obligation.
- There are 1,731 unique product service codes currently in the data, these product service codes were rolled up into 188 parent product service codes that reflect the general category of the product or service being provided.

3. FEDERAL ACCOUNTS DENDROGRAM



- Endpoint used: Awards
- We used the fields:
    - treasury_account_identifier.agency_id
    - treasury_account_identifier.federal_account.main_account_code
    - treasury_account_identifier.funding_toptier_agency.fpds_code
    - obligations_incurred_total_by_tas_cpe
    - treasury_account_identifier.federal_account.id, and
    - treasury_account_identifier.federal_account.account_title
- These fields were grouped and aggregated the obligations data based on the Treasury Account Number (Agency ID plus main account code), then merged the endpoint data with the OMB hierarchy of Agency and Bureau Titles and corresponding Treasury Account Numbers for all active accounts. For agencies where the agency id was more than 2 digits long, we used the first 2 digits of the FPDS code for the agency id instead. This merge allowed us to add obligation amounts to the hierarchy of federal accounts and then display this information through the dendrogram's tree structure.
- We then made a few changes to make the information easier to navigate, including grouping all non-CFO act agencies together and pulling out or grouping together offices when the number of levels within an agency made it difficult to display through the dendrogram's set structure. In cases where a level within the agency would have been excluded, we added it at the end of the account name to be more descriptive. The federal account nodes link to the federal account pages on USAspending.gov based on the account Id.

Data Dictionary

Federal financial reporting elements are not always intuitive. For the official definitions of each data element you will find

The latest Data Dictionary is linked on the “Data Model” section of the DATA Act GitHub page.

API Guide

This guide provides users of the USASpending API with information to facilitate use of data accessible from the API and lower the barriers to entry for new users.

A list of the endpoints and short descriptions of each can be found at the USASpending API website.

API Endpoints

The API provides several endpoints, of those a summary is provided for the data that is available to export from 8 key endpoints using a get request, as well as the endpoint address.

A. **Accounts Awards -** /api/v1/accounts/awards/

Financial account data by TAS account includes program activity, award id and object class. Each observation will include the relevant identifiers for PIID, FAIN, URI, treasury account data, awarding and funding agency, and federal account budget accounts data.

B. **Awards** - /api/v1/awards/

List of award records. Each observation identified by "id", includes type of award, Procurement Instrument Identifier (PIID), Federal Account identification Number (FAIN), Unique Record Identifier (URI)[1], amount, awarding and funding agency details, as well as recipient details and place of performance information.

C. **Federal Accounts** - /api/v1/federal_accounts/

List of Federal Accounts. Observations are identified by "id", and includes the following: agency identifier, main account code, account title, and the concatenated federal account code.

D. **Treasury Appropriations Accounts (TAS) -** /api/v1/tas/

List of treasury appropriation accounts by Treasury Account Symbol (TAS). Observations are identified by the treasury account identifier, includes federal account information, reporting, funding, and awarding agency information as well as budget function codes.

A note on "TAS"…

TAS is a complex concept, so let's break this down each component for clarity. Each TAS is composed of:

- **Main Account Code** (MAC): 4 digits, Forms the bases of the federal account
- **Sub Account Code** (SAC): 3 digits, Rare, used in giant agencies that need to split up their accounts. Default is 000
- **Agency ID**: Agency Identifier, assigned by congress. Same role as CGAC
- **Allocation Transfer Agency ID** (ATA): Allocation Transfer Agency ID (usually not included)
- **Beginning period of availability** (BPOA): year when obligations are s available **AND**
- **Ending Period of availability** (EPOA): when they can no longer give obligations
  OR
- **ATC:** Instead of specific years, an "X" represents an indefinite account: it can be used until deauthorized

E. **TAS Balances -** /api/v1/tas/balances/

List of treasury appropriation account balances by quarter and by TAS. Observations identified by appropriation account balance id, includes Current Period End (CPE) and Fiscal Year End (FYE) obligations and outlay data, as well as treasury account, budget function codes, federal account, awarding and funding agency information.

[1] URI is only used where the combination of an assistance observation is not unique to a FAIN and Modification number. URI is defined by the agency.

F. **TAS Categories** - /api/v1/tas/categories/

List of treasury appropriation account balances by fiscal year, TAS, program activity, and object class. Observations are identified by field named "financial account by program activity object class id", include obligations and outlay information by CPE and FYE, program activity, submission, object class, treasury account identifiers, budget function, and funding and awarding agency identifiers.

G. **Subawards** - /api/v1/subawards/

List of subaward records. Observations identified by "id", include subaward number; award data; recipient details; North American Industry Classification System (NAICS) codes; submission identifiers; awarding, funding, and subtier awarding agency identifiers.

A note on "Obligations" v "Outlays"
Obligations are promises to pay for goods or services at a later point in time. This does not mean money spent as agencies still have this money. Outlays, on the other hand, is money spent and gone.

H. **Transactions -** /api/v1/transactions/

List of transactions including contracts, grants, loans, etc. Observations are identified by "id", includes transaction description, amount and date, as well as awarding, funding, and subtier agency engaging in the transaction. Transactions also include detailed recipient, contract, and place of performance data.

Table 1: Endpoint Unique Identifiers includes each endpoint and the identifiers that can be used to identify unique observations within each dataset.

Table 1: Endpoint Unique Identifiers

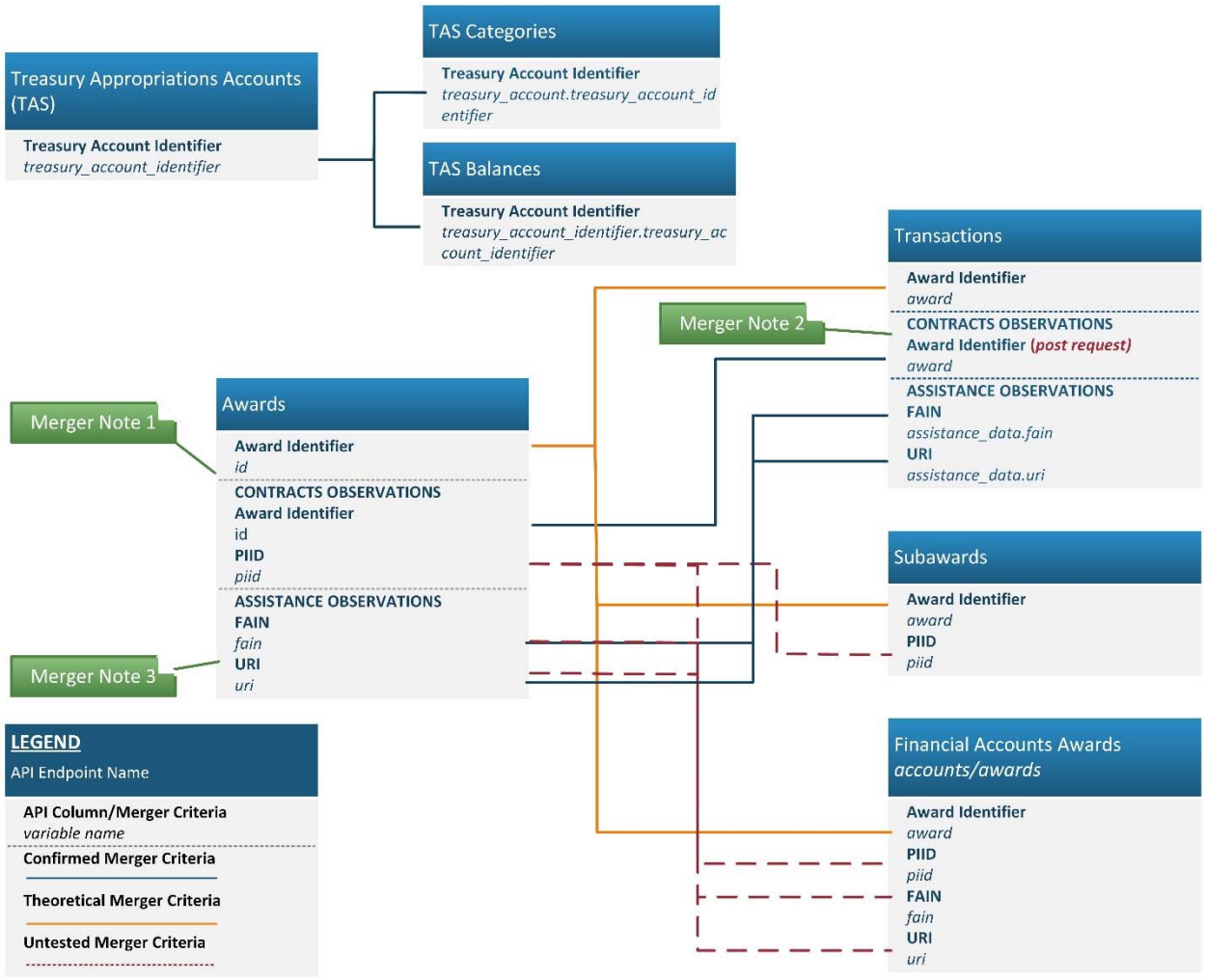
Endpoint Title	Unique Identifier
Accounts Awards	financial_accounts_by_awards_id
Awards	id
Federal Accounts	No unique identifier available use federal_account_code
Treasury Appropriation Accounts (TAS)	treasury_account_identifier tas_rendering_label
TAS Balances	appropriation_account_balances_id; or Combine: treasury_account_identifier.treasury_account_identifier[2] treasury_account_identifier.tas_rendering_label[3]
TAS Categories	financial_accounts_by_program_activity_object_class; or Combine: treasury_account_identifier.treasury_account_identifier; program_activity.id; object_class.object_class
Subawards	Not available
Transactions	id

[2] This will require some cleaning due to duplication of data submitted more than once for the same account but should be unique when duplicates are identified and removed.
[3] This will require some cleaning due to duplication of data submitted more than once for the same account but should be unique when duplicates are identified and removed.

Merging Endpoints

For the endpoints listed above Figure 1: Endpoint Merging Map visualizes criteria that can be used to connect relevant datasets. This is a living document and will be updated as merger criteria is tested and confirmed.

Figure 1: Endpoint Merging Map



1. **Merger Note 1:** The Awards dataset generated by the get request can only be merged by separating observations into two groups, Contracts and Assistance observations, and merging by the criteria outlined above for each respective group.
2. **Merger Note 2:** The Awards dataset and transactions dataset theoretically should be able to be merged on id and award, respectively. However, this will require the use of the transactions post request to include the award column, which is not available in the get request.
3. **Merger Note 3:** Matching Award assistance observations suffers from data quality issues, therefore merging all assistance entries first on FAIN, then creating a second merge dataset merged on URI results in the greatest number of retained records. Note, to combine into one assistance dataset, each merged file will need to have the same columns (remove the unmerged duplicate FAIN/URI column from each), appended the datasets, then remove duplicates from the final complete assistance dataset.

Table 2: Endpoint Merger Criteria

Dataset A	Data Set and Merge Variables
Awards	**Transactions:** • Contract Entries**:** o id – award (only available in the transactions post request) • Assistance Entries: o uri – assistance_data.uri o fain – assistance_data.fain **Accounts Awards** • Theoretically a complete match should be available using: o award to id (creates approx.. 30% match on awards) • <u>Three-part merger:</u> o First match all observations, and drop all unmatched observations: ▪ id to award o For Contract Entries, merge using: ▪ piid o Assistance Entries (direct payments, loans, insurance, grants), merge using: ▪ Fain and uri o Retain only matched for each merger, and then rbind the matched observations. Lastly delete all duplicated entries.
TAS Balances	**TAS Categories:** • treasury_account_identifier.treasury_account_identifier to treasury_account_identifier.treasury_account_identifier

Geographic Data

Geographic data is available for within the following endpoints awards, transactions, and subawards. Geographic data can be inconsistently reported for therefore an inventory and recommended use of the information available (variables) within the data is summarized below. The subawards endpoint is not currently available therefore an evaluation of subawards geographic data is not included below.

Percentages reported below are based on data reported through the third quarter submissions for fiscal year 2017. As additional data becomes available counts and recommendations will be updated.

Awards Geographic Data Available

Endpoint: /api/v1/awards/

Geographic data is provided at two levels within the awards data set, address, city, state and zip can be entered for the award recipient and for the place of performance.

Recommended Use

Recipient state-level data is the most complete and reliable geographic data available. Use of the recipient's street address, city, and state will provide more granular data but will reduce the reliability and the completeness of the sample of data available. There is a notably less information reported for foreign recipients. The most granular level of data recommend for use for place of performance is state and country for domestic places of performance. To include places of performance world-wide the most granular data would be at the country-level, due to the low rate of reporting.

Table 3: Award Data Recommended Geographic Data

Geographic Data	Variable Names	Percent Reported
Recipient Data		
• City	recipient.location.city_name	64%
• State	recipient.location.state_code recipient.location.state_name	74% 74%
• Country	recipient.location.country_name recipient.location.country_code	100% 100%
Place of Performance Data		
• State	place_of_performance.state_code place_of_performance.state_name	85% 85%
• Country	place_of_performance.country_name place_of_performance.country_code	97% 97%

Recipient Geographic Data

Of the recipient data that was reported, 99 percent of recipients were located in the United States, and 1 percent of recipients reported a foreign location.

Table 4: Awards Recipient Geographic Data Inventory

Recipient Geographic Data	Variable Name(s)	Percentage Reported
Street Address[4]	recipient.location.address_line1 recipient.location.address_line2 recipient.location.address_line3	62% 5% 1%
City[5]	recipient.location.city_name	64%
State[6]	recipient.location.state_code recipient.location.state_name recipient.location.state_description	74% 74% 0%
5 Digit Zip Code[7]	recipient.location.zip5	24%
Country	recipient.location.country_name recipient.location.country_code	100% 100%
Foreign - Street Address[8]	recipient.location.address_line1 recipient.location.address_line2 recipient.location.address_line3[9]	99% 11% 0%
Foreign - City[10]	recipient.location.foreign_city_name recipient.location.city_name	1% 51%
Foreign - Province[11]	recipient.location.foreign_province	1%
Foreign - Postal Code[12]	recipient.location.foreign_postal_code	1%

Data Quality Concerns:

- State abbreviations and Names
  - State code and name variables do not have a character limits or validations in place therefore there are instances where full state or province names are used rather than the code and vice versa. For example, “CA” and “California” are both present in state name column.
  - State name and code variable includes foreign countries names which were incorrectly reported in that field and should be moved to the country field and domestic state variables left blank for foreign recipient or place of performance observations.
  - Domestic state names and abbreviations are included for foreign addresses, potentially due to reporting in the default field rather than the specified foreign address/ foreign city fields.
- Street Addresses:
  - Over 2,900 addresses where entered as “ANY VALUE”.

[4] Percentage of data reported for recipients that reported located within the USA.
[5] Evaluated the percent of recipients that indicated they were located within the USA.
[6] Evaluated the percent of recipients that indicated they were located within the USA.
[7] Evaluated the percent of recipients that indicated they were located within the USA.
[8] Evaluated the percent of recipients that indicated they were located outside the USA.
[9] Recipients that indicated they were located outside the USA but that reported a city name in the default recipient city name field.
[10] Evaluated the percent of recipients that indicated they were located outside the USA.
[11] Evaluated the percent of recipients that indicated they were located outside the USA.
[12] Evaluated the percent of recipients that indicated they were located outside the USA.

  - The ability to verify the validity of individual street addresses is limited and these entries should be used with caution.
- Zip codes are not validated, therefore there may be entries that that do not meet the 5 digit string of numbers requirement once all other characters are excluded.
- Some address information must be suppressed to preserve personally identifiable information.

Place of Performance Geographic Data

Place of performance data is not as consistently reported as the recipient information. Additionally, foreign entries will provide little detail beyond the country reported.

Table 5: Awards Place of Performance Geographic Data Inventory

Place of Performance Geographic Data	Variable Name(s)	Percentage Reported
Street Address[13]	place_of_performance.address_line1 place_of_performance.address_line2 place_of_performance.address_line3	0% 0% 0%
City[14]	place_of_performance.city_name	18%
State[15]	place_of_performance.state_code place_of_performance.state_name place_of_performance.state_description	85% 85% 0%
5 Digit Zip Code[16]	place_of_performance.zip5	0%
Country	place_of_performance.country_name place_of_performance.country_code	97% 97%
Foreign - City[17]	place_of_performance.foreign_city_name	0%
Foreign - Province[18]	place_of_performance.foreign_province	0%
Foreign - Postal Code[19]	place_of_performance.foreign_postal_code	0%

Data Quality Concerns:

- Place of performance data for observations occurring outside of the United States are not reliable, with almost all missing values for foreign zip code, state, city, and province.

Transactions Geographic Data Available

Endpoint: /api/v1/transactions/

Geographic data is provided at two levels within the transactions data set, address, city, state and zip can be entered for the recipient and for the place of performance.

[13] Evaluated the percent of recipients that indicated they were located within the USA.
[14] Evaluated the percent of recipients that indicated they were located within the USA.
[15] Evaluated the percent of recipients that indicated they were located within the USA.
[16] Evaluated the percent of recipients that indicated they were located within the USA.
[17] Percentage of observations that were reported as a place of performance located outside the USA.
[18] Percentage of observations that were reported as a place of performance located outside the USA.
[19] Percentage of observations that were reported as a place of performance located outside the USA.

Recommended Use

Geographic data for transactions reported consistently for domestic recipient locations. Use of recipient state (state code) and country is the most complete. More granular data is available for recipients with street address for recipients being reported for approximately 70% of recipients. Domestic transactions report data more consistently, foreign transactions for both recipient and place of performance show a low rate of reporting.

Table 6: Recommended Geographic Data for Transactions

Geographic Data	Variable Names	Percent Reported
Recipient Data		
• City	recipient.location.city_name	73%
• State	recipient.location.state_name recipient.location.state_code	77% 77%
• Country	recipient.location.country_name recipient.location.country_code	100% 100%
Place of Performance Data		
• State	place_of_performance.state_code	87%
• Country	place_of_performance.country_name place_of_performance.country_code	96% 96%

Recipient Geographic Data

Domestic recipient data is reported consistently for first address line, city, state, and country. Foreign recipient data is not well reported with only recipient country reported consistently. Recipient country is missing for over 3,400 transactions.

Table 7: Transactions Recipient Geographic Data Rate of Reporting

Recipient Location Geographic Data	Variable Name(s)	Percentage Reported
Street Address	recipient.location.address_line1 recipient.location.address_line2 recipient.location.address_line3	73% 5% 1%
City	recipient.location.city_name	73%
State[20]	recipient.location.state_name recipient.location.state_code recipient.location.state_description	77% 77% 0%
Zip Code[21]	recipient.location.zip5	37%
Country[22]	recipient.location.country_name recipient.location.location_country_code	100% 100%
Foreign – City	recipient.location.foreign_city_name	1%
Foreign - Province	recipient.location.foreign_province	38%
Foreign – Postal Code	recipient.location.foreign_postal_code	1%

[20] Percentage of recipient data reported for each variable calculated using only domestic observations.
[21] Percentage of recipient data reported for each variable calculated using only domestic observations.
[22] Recipient country is not reported for 76 observations.

Place of Performance Geographic Data

Place of performance data is most reliable for state and country, using state code which is reported more reliably than state name.

Table 8: Transactions Place of Performance Geographic Data Rate of Reporting

Place of Performance Geographic Data	Variable Name(s)	Percentage Reported
Street Address	place_of_performance.address_line1 place_of_performance.address_line2 place_of_performance.address_line3	0% 0% 0%
City	place_of_performance.city_name	33%
State[23]	place_of_performance.state_name place_of_performance.state_code place_of_performance.state_description	87% 87% 0%
Zip Code[24]	place_of_performance.zip5	0%
Country	place_of_performance.country_name place_of_performance.location_country_code	96% 96%
Foreign – City	place_of_performance.foreign_city_name	0%
Foreign - Province	place_of_performance.foreign_province	0%
Foreign – Postal Code	place_of_performance.foreign_postal_code	0%
Foreign - Description	place_of_performance.foreign_location_description	0%

- Data Quality Concerns:
  - Place of performance state name includes many invalid entries including state code abbreviations and foreign country names.
  - Recipient state name and code may include US territories as a state and include some invalid entries such as “Other” and “Curacao”.
  - Many zip codes include less than 5 digits, this could be a problem generated by importing the zip code values as a number rather than a factor or string of characters to preserve zeros.
  - Place of performance country is not reported for 4 percent of observations.

[23] Percentage of place of performance data reported for each variable calculated using only domestic observations.
[24] Percentage of place of performance data reported for each variable calculated using only domestic observations.