Open Data Automation Guidelines

This document provides guidance on several automation options for data publication to the Connecticut Open Data Portal (data.ct.gov). This guidance is intended for Agency Data Officers and other publishers of open data within Connecticut state government. For assistance or more information about these options, contact the Open Data Portal Administrators at OPM via dapa@ct.gov.

Overview

As traffic continues to grow on the Open Data Portal, so does the need to keep datasets up to date. Agencies may find themselves with tens or even hundreds of datasets that need to be updated on a scheduled basis in order to keep data current and within the open data portal publication and retirement guidelines. Rather than updating these datasets manually, a more time- and resource-efficient option is to update them automatically.

This document will cover a range of different automation options for the Open Data Portal. The automation tools covered include:

- DataSync,
- PilotFish,
- Socrata Gateway,
- Socrata-py (Python), and
- RSocrata (R).

These tools are evaluated on a set of criteria, including the download, set-up, first time use, and regular use difficulty of the automation tool. The criteria are represented by two measures: set-up difficulty (this covers download and set-up) and use difficulty (this covers first time and regular use). Both measures are on a scale from 1 to 5 where 1 requires the least amount of difficulty and 5 the most amount of difficulty. Difficulty ratings are subjective and intended only as a guide. Further information is provided for each tool including resources on where to download these tools from as well as step-by-step directions for preliminary actions and set-up. Multiple tables are also available at the end of the document, which allow for easy comparison of technical aspects within some of the tools.

NOTE: Considering all these tools require an owner/publisher Socrata account, agencies may want to use accounts that correspond to a team or agency rather than an individual to avoid dependance on a single individual's account. This will allow for multiple people within an agency to access and maintain data set automation, regardless of which user's account is associated with the Socrata connection/access for these tools. To request a Socrata account for your team, contact the Open Data Portal Administrators at OPM via dapa@ct.gov.

DataSync

Set-up Difficulty: 2 | Use Difficulty: 2

What it is

DataSync is an executable Java application that provides replace, append, delete, and upsert capabilities on a dataset. This tool allows for automated updates when used alongside a task scheduler such as Windows Task Scheduler.

How it works

Users can access DataSync processes through the graphical user interface (GUI) or in the command line. (This guideline only covers using the GUI.) DataSync allows for CSV or TSV file uploads from your local machine or a networked hard drive such as SharePoint. Once the user has set up their DataSync task, it can be saved as a .sij file. This file can then be used in a basic task in an application like Windows Task Scheduler.

Prerequisites

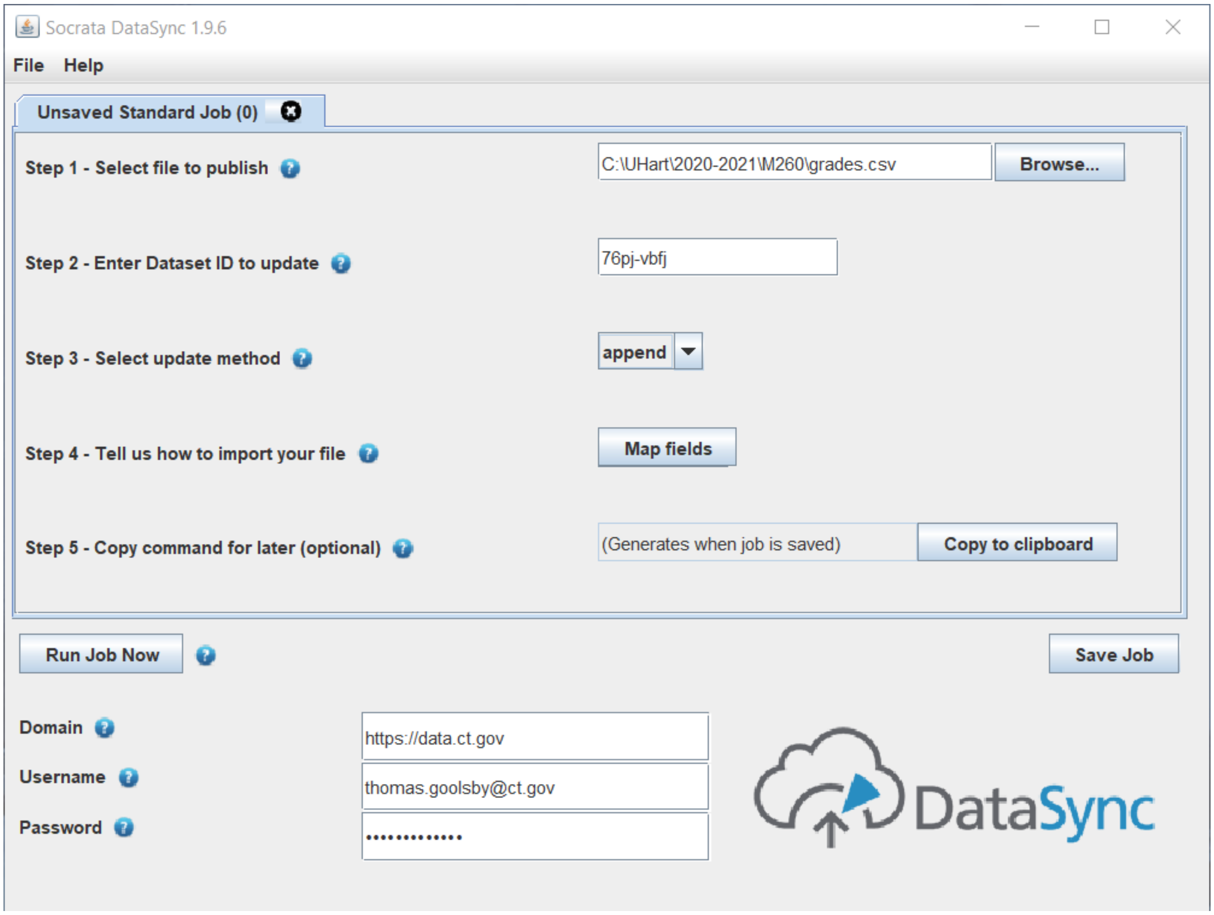
- Computer or server running Java 8 or higher
  - The latest version of Java can be downloaded here.
- Socrata account with the publisher or owner role
- An app token for the desired dataset. (The app token will be the Socrata Dataset ID of your dataset. This can be found by accessing the Socrata API from the Open Data Portal. The ID will be in the form of xxxx-xxxx.)

When is this tool a good choice?

DataSync is a great tool for someone who is looking to simply update a dataset. Furthermore, the set-up for this tool is very simple when compared to other tools in this document. One of the biggest issues, however, is that for this tool to automatically update, it must be run through a task scheduler. Windows Task Scheduler, for example, will not run tasks if the computer is not on, so this tool requires some level of user responsibility, even after set-up. For more information and to download DataSync, please visit the Socrata website here.

Process Overview

1. Select the file from your local hard drive that you wish to upload to your data set. This must be a CSV or TSV file.
2. Enter the Socrata Dataset ID of your dataset. This can be found by accessing the Socrata API from the Open Data Portal. The ID will be in the form of xxxx-xxxx.
3. Choose from one of four update methods: replace, append, upsert, or delete.
   - Replace will replace the current dataset on the Open Data Portal with the dataset from Step 1.
   - Append will add all rows from the dataset in Step 1 to the end of the dataset on the Open Data Portal.
   - Upsert will update the current dataset on the Open Data Portal as well as append any new rows from the dataset in Step 1.
     - IMPORTANT: A Row Identifier must be in place within select columns of the dataset on the Open Data Portal. For more information about how to set up a row identifier on the Open Data Portal, refer to the Socrata support article Setting a Row Identifier in the Socrata Data Management Experience.
   - Delete will remove all rows in the dataset on the Open Data Portal that match the Row Identifiers from the dataset in Step 1. Essentially, the dataset from Step 1 would be a single column containing the Row Identifiers to be deleted.
     - As with the Upsert option, the Delete option also requires Row Identifiers to be in place in the dataset on the Open Data Portal.
4. Click the “Map fields” button to check that the columns from your dataset in Step 1 map to the correct columns in the dataset on the Open Data Portal.
5. Enter in the required information (domain, username, password) in the fields at the bottom. The domain will be https://data.ct.gov, and the username is your email for accessing the Open Data Portal and the password is your Open Data Portal password.



Step 1 - Select file to publish: C:\UHart\2020-2021\M260\grades.csv

Step 2 - Enter Dataset ID to update: 76pj-vbfj

Step 3 - Select update method: append

Step 4 - Tell us how to import your file

Step 5 - Copy command for later (optional): (Generates when job is saved)

Domain: https://data.ct.gov

Username: thomas.goolsby@ct.gov

Password: •••••••••••••

Screenshot of the DataSync GUI

Pilotfish

Set-up Difficulty: 2 | Use Difficulty: 5

What it is

PilotFish is an extremely powerful application tool that combines the PilotFish IDE with the XCS eiConsole. When put together, PilotFish capabilities can be operated through the XCS eiPlatform desktop application. Scheduled updates can be run with ease since PilotFish has a built-in scheduler with a wide range of timing options.

How it works

PilotFish works as a pipeline from the source location of the user's file to a Socrata server, which then connects to the Open Data Portal. The user first creates a new route, where they can choose what kind of source file they wish to reference. The most used source locations are Excel files, CSV files, or a database. This route then takes the source file, performs source transformations, performs target transformations, and then delivers the transformed file to the target location, which would be the Open Data Portal in this situation. For complex transformations, the use can create transformation rules for proper data integration to the Open Data Portal.

Prerequisites

- Socrata account with the publisher or owner role
- An app token for the desired dataset. (This can be found on the Socrata API.)

When is this tool a good choice?

If a data set needs advanced transformations to go from a source file to the Open Data Portal, then PilotFish would be a great tool for that job. However, since many of the update jobs on the Open Data Portal simply require an append or replace function, less complex tools may be a better fit. Moreover, PilotFish jobs can only be accessed through the computer in which they were created on. If there was an issue with the user's computer, then the PilotFish route would have to be recreated on another user's computer. For this reason, DAS recommends installing PilotFish on an application server, which is almost always up, and are monitored, baked up, and can be given direct access to backend databases.

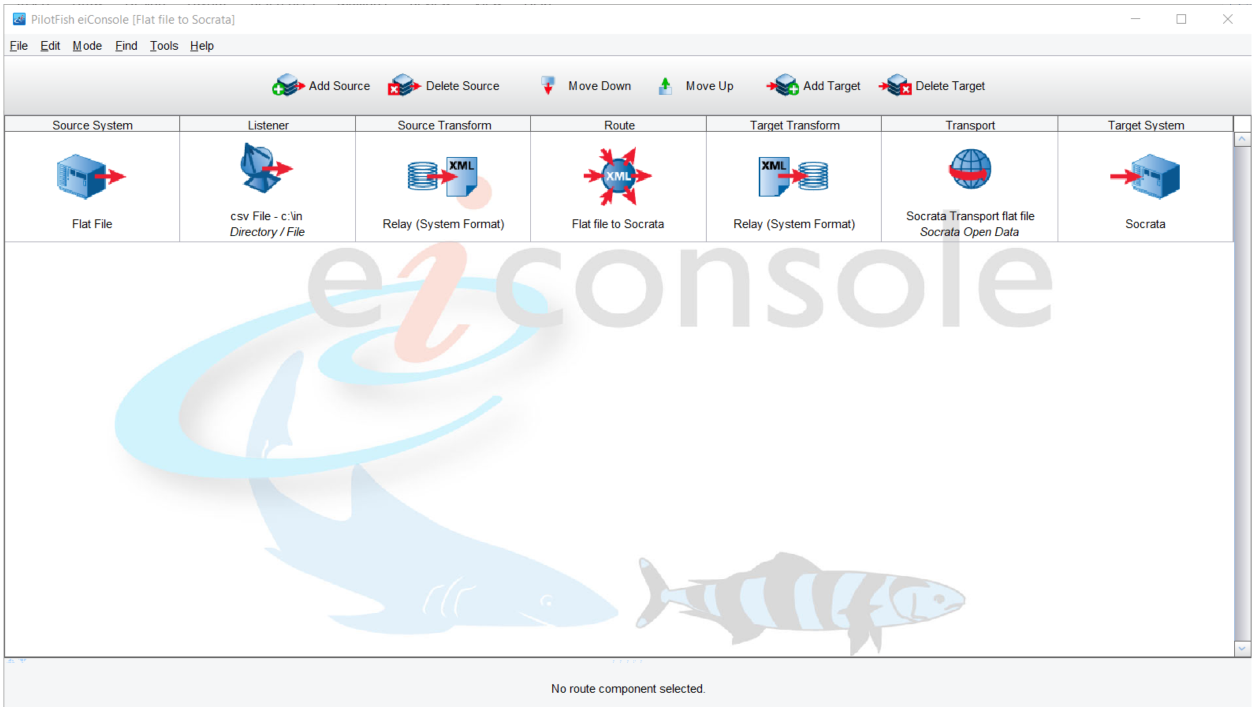
If any agency did wish to use PilotFish, they could get access easily since the State of Connecticut has a license for PilotFish services. For more information about the use of PilotFish by Connecticut state agencies, view the enterprise service agreement, available here.

For more information on how to set-up and use PilotFish, please contact Matt Shea from the Department of Administrative Services, at matt.shea@ct.gov.

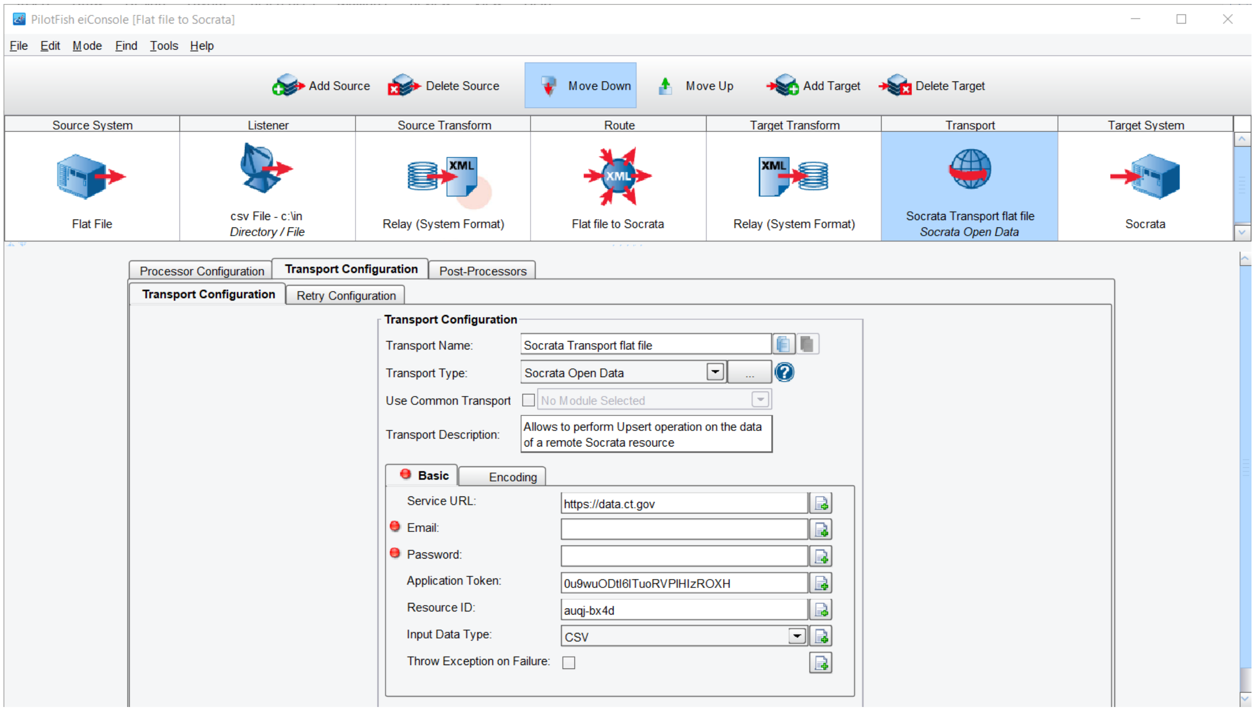
To download PilotFish, visit the following link and click the red "Free Trial" button and log in with your ct.gov email address: https://www.pilotfishtechnology.com/. Those who sign up for the free trial with a valid ct.gov email address should automatically receive a license key.

Process Overview

Once you have PilotFish downloaded, visit the following link for an extensive collection of documentation provided by PilotFish here.



Basic PilotFish route from flat file (CSV) to data set on the Open Data Portal



Transport Configuration

Transport Name: Socrata Transport flat file

Transport Type: Socrata Open Data

Use Common Transport □ No Module Selected

Transport Description: Allows to perform Upsert operation on the data of a remote Socrata resource

Service URL: https://data.ct.gov

Email:

Password:

Application Token: 0u9wuODtl6ITuoRVPlHIzROXH

Resource ID: auqj-bx4d

Input Data Type: CSV

Throw Exception on Failure: □

Required information to create connection to data set on the Open Data Portal

Gateway

Set-up Difficulty: 4 | Use Difficulty: 1

What it is

Socrata Gateway is a tool that has been directly integrated into the Socrata Dataset Management Experience (SDMX). This tool is set up and used within the Open Data Portal itself which allows for a one-time set up for scheduled updates.

How it works

Gateway is used on the Open Data Portal and can be set up for already existing datasets as well as new ones. A server connection, called an Agent, is downloaded by the user onto the server or computer they are connecting to Socrata. Essentially, this Agent sets up a connection between the dataset on Socrata and the file directory on the user's server or computer. Once this Agent connection is set up and working, the user can simply use Plug-ins to choose what types of files will be uploaded/updated on the Open Data Portal. Some useful plug-ins include CSV, TSV, and Excel; these are likely the most common plug-ins that are necessary for most datasets on the Open Data Portal. Unlike other tools, Gateway has a scheduler built in, meaning that daily, weekly, monthly, or annual updates can be scheduled. The scheduler allows the user to enter a specific time for updates as well. Once the scheduled update time is determined and entered, the user will have automatic updates set up that will perform a replace on the data.

Prerequisites

- Socrata account with the publisher or owner role

When is this tool a good choice?

Socrata Gateway is one of the best choices in this guideline document for just about every data set on the Open Data Portal. Although the set-up difficulty and time are higher than other tools, the ease of use heavily outweighs that set-up difficulty. Since Gateway is directly integrated into the Socrata framework used by the Open Data Portal, setting scheduled updates is extremely easy and intuitive. For most, the set-up time will take around 30 minutes or so However, this set-up is a one-time process for each data set you wish to automate. If the Open Data Portal continues to use Socrata services, Gateway is likely to be the most up-to-date and seamless option for automation.

A log of known issues with Gateway is available on the Socrata website here.

Process Overview

To set up a Gateway connection for a CSV based dataset on the Open Data Portal, follow the instructions below. To access the Gateway page on the Open Data Portal for a:

New dataset

1. Once logged onto the Open Data Portal, click the Create tab on the menu bar at the top of the screen.
2. Select the Dataset option.
3. Enter the name of your dataset.
4. Select the Add Data option on the homepage of your new dataset.
5. Select the third option down on the menu at the left of the screen – Connect to an External Data Source (Socrata Gateway).
6. You are now ready to set up your Gateway connection for this dataset.

Existing dataset

1. Once logged onto the Open Data Portal, go to the homepage of the dataset for which you would like to setup a Gateway connection.
2. Click the Edit button near the top right of the homepage.
3. Select the Review & Configure Data option on the homepage of your dataset.
4. At the top left of the screen, click the Choose Data Source link – At the top left of the screen it will look like: Choose Data Source > Preview.
5. Select the last option on the menu at the left of the screen – Connect to an External Data Source (Socrata Gateway).
6. You are now ready to set up your Gateway connection for this dataset.

To create a new agent and plugin for your Gateway connection follow the steps below:

1. Once on the Gateway page for your dataset, select the Provision new agent option at the top right of the screen.
2. Fill out the Agent Name field.
   a. Suggested convention for naming the agent: Name of the person who is responsible for the agent + The data source type (e.g. Zaldonis_CSV).
3. After you have named the agent, click the Download Agent button.
4. Allow the agent to download.
5. Select the agent and extract/unzip the compressed folder to the location of your choice.
6. FOR WINDOWS – Open the unzipped agent folder and right click the install.bat file (Windows batch file).
7. Select Run as Administrator.

8. The command prompt will open; give a name to your service, selecting a name that makes sense to you.
9. Press enter to submit your service name and press any key to close the command prompt once the job is complete.
10. Go back to your browser tab on the Open Data Portal and refresh the "Am I Connected?" box at the bottom of the Provision Agent pop-up.
11. Your agent should now be connected.
12. Click the Next button.
13. Select the Set-up a plugin bubble and click the Next button.
14. On the next page, you should see the CSV plugin; click the Set-up button.
15. After reading the overview page for the CSV plugin, click the Next button.
16. Fill out the Plugin Name field.
    a. Suggested convention for naming the plugin: Name of the person who is responsible for the agent + The data source type.
17. Click the Next button.
18. Follow the Set-up instructions on the next screen – this requires you to locate your data path of you agent as well as use the command prompt to run a single command.
19. After the command has been run, a new window will open, once here, simply select the root directory for all CSV files that are accessed by your dataset.
20. Your Gateway agent and plugin should now be connected!

For additional resources on setting up Socrata Gateway, refer to the following Socrata resources:

1. Gateway Overview
2. Socrata Gateway Technical Overview and Requirements
3. Socrata Gateway for Data Publishers

Socrata-py (Python)

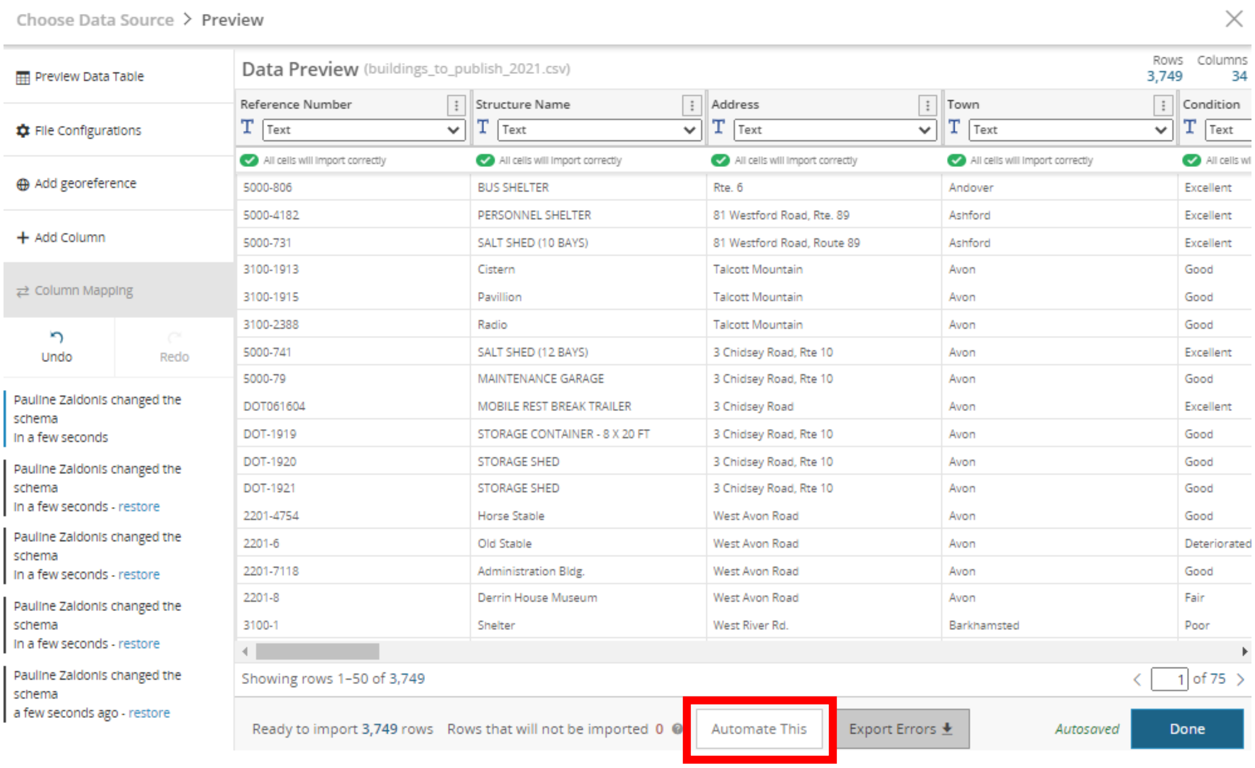
Set-up Difficulty: 4 | Use Difficulty: 2

What it is

The Socrata-py package provides Socrata methods and automation capabilities using the Python language. Like Gateway, there is some integration of Socrata-py into the SDMX.

How it works

Socrata-py is a Socrata generated Python file that the user can run on a scheduled basis using a task scheduler such as Windows Task Scheduler. Before accessing the Python file, the user must set up environment variables on their local machine for their Socrata username and password. Once this preliminary step is taken, the user can simply create their new dataset on the Open Data Portal and click the "Automate This" button in the Data Preview window.



Data Preview (buildings_to_publish_2021.csv)

Rows 3,749 Columns 34

Reference Number	Structure Name	Address	Town	Condition
Text	Text	Text	Text	Text
All cells will import correctly	All cells will import correctly	All cells will import correctly	All cells will import correctly	All cells wi
5000-806	BUS SHELTER	Rte. 6	Andover	Excellent
5000-4182	PERSONNEL SHELTER	81 Westford Road, Rte. 89	Ashford	Excellent
5000-731	SALT SHED (10 BAYS)	81 Westford Road, Route 89	Ashford	Excellent
3100-1913	Cistern	Talcott Mountain	Avon	Good
3100-1915	Pavillion	Talcott Mountain	Avon	Good
3100-2388	Radio	Talcott Mountain	Avon	Good
5000-741	SALT SHED (12 BAYS)	3 Chidsey Road, Rte 10	Avon	Excellent
5000-79	MAINTENANCE GARAGE	3 Chidsey Road, Rte 10	Avon	Good
DOT061604	MOBILE REST BREAK TRAILER	3 Chidsey Road	Avon	Excellent
DOT-1919	STORAGE CONTAINER - 8 X 20 FT	3 Chidsey Road, Rte 10	Avon	Good
DOT-1920	STORAGE SHED	3 Chidsey Road, Rte 10	Avon	Good
DOT-1921	STORAGE SHED	3 Chidsey Road, Rte 10	Avon	Good
2201-4754	Horse Stable	West Avon Road	Avon	Good
2201-6	Old Stable	West Avon Road	Avon	Deteriorated
2201-7118	Administration Bldg.	West Avon Road	Avon	Good
2201-8	Derrin House Museum	West Avon Road	Avon	Fair
3100-1	Shelter	West River Rd.	Barkhamsted	Poor

Showing rows 1–50 of 3,749

1 of 75

Ready to import 3,749 rows Rows that will not be imported 0

Pauline Zaldonis changed the schema
In a few seconds

Pauline Zaldonis changed the schema
In a few seconds - restore

Pauline Zaldonis changed the schema
In a few seconds - restore

Pauline Zaldonis changed the schema
In a few seconds - restore

Pauline Zaldonis changed the schema
a few seconds ago - restore

Socrata provides update and replace capabilities, which will generate a unique Python script for your dataset and the chosen update function (update or replace). The user can then copy this code from the Open Data Portal, paste it into a text editor (such as Notepad), and save the file using the Python (py) extension.

Prerequisites

- Socrata account with the publisher or owner role
- Computer running Python3 or higher
- The latest version of Python can be downloaded here.
- Environment variables for Socrata username and password on your local machine

When is this tool a good choice?

The Socrata-py package is an excellent choice for many datasets on the Open Data Portal. Since there is direct integration into the SDMX, there is minimal set-up needed outside of the Open Data Portal. This guideline document provides a step-by-step of how to get those environment variables set, however, this process still requires knowledge of navigating some advanced settings within your local machine. Furthermore, this tool relies on a task scheduler to be run automatically, thus automatic updates won't be run unless your computer is on. Although this tool is essentially a Python script, the user does not need any knowledge of the Python programming language. Once the Python script is generated from the Open Data Portal, there is no coding needed going forward. For more information, please visit this page on the Socrata website.

Process Overview

How to set up required environment variables on your local machine FOR WINDOWS:

1. Click the Windows icon in the bottom left of the screen and go to your Settings.

2. Select the System option and then select the About option at the bottom of the menu at the left of the screen.
3. Scroll down to the bottom of the screen and select the Advanced system settings option.
4. A new window will open, and on the Advanced tab, click the Environment Variables button.
5. Click the New button under the “User variables for …” list.
6. Set the variable name to MY_SOCRATA_USERNAME and the value as your ct.gov email used to log into the Open Data Portal and click OK.
7. Click the New button again and this time, set the variable name to MY_SOCRATA_PASSWORD and the value as your Open Data Portal password, then click OK.
8. Click OK again on the Environment Variables window, and one more time on the System Properties window.
9. To check that your environment variables are set up correctly, open the Command Prompt.
10. Enter echo %MY_SOCRATA_USERNAME% and your ct.gov email should be returned.
11. Enter echo %MY_SOCRATA_PASSWORD% and your Open Data Portal password should be returned.
12. Your environment variables are all set up!

How to automate a data set using Python scripts:

1. Once logged onto the Open Data Portal, click the Create tab on the menu bar at the top of the screen.
2. Select the Dataset option.
3. Enter the name of your dataset.
4. Select the Add Data option on the homepage of your new dataset.
5. Select the first option on the menu at the left of the screen – Upload a data file.
6. Locate your data file and load the data into the data preview.
7. At the bottom of the preview page, click the Automate This button.
8. Choose if you would like to run an update or replace transformation on the data set.

    Update will append any rows in the data file to the data set on the Open Data Portal.

    Replace on the other hand will replace the data set on the Open Data Portal with the data file on your local machine.
9. Follow the instructions on the screen, which include some lines of Python code that can be copy and pasted.
10. Install the socrata-py Python library by going into your local Python environment, such as the command prompt, and running this line of code: pip install socrata-py~=1.0.0. You may want to check to see if you have Pip installed by running: pip --version
11. Create a .py file containing the chunk of code on the screen starting with the line, “from socrata.authorization import Authorization”. Name it something you will recognize (ex. my-update-script.py), and save the file to the same location as the data file you uploaded to the Open Data Portal.
12. Go back to the Open Data Portal, fill out any required metadata, and upload the data set.
13. Once the data is uploaded to the Open Data Portal, go to your local Python environment, and run your Python script containing the code from Step 11 to make sure that all environment variables are set up correctly.
14. Open a task scheduler, such as Cron or Windows Task Scheduler.
15. Create a new basic task, name it, and set the frequency and run time for your task.

16. For the action of the task, put this: python3 [your Python script name here].
17. Save the task and now your data set is all set up to be updated automatically using Python!

RSocrata (R)

Set-up Difficulty: 2 | Use Difficulty: 3

What it is

RSocrata is a package developed by the City of Chicago that allows the user to pull data sets from the Open Data Portal as well as write R data frames to the Open Data Portal. This tool does require some familiarity with the R programming language.

How it works

The RSocrata package contains many functions and capabilities, two of which are almost always necessary when using this tool for data set automation. The read.socrata function lets the user read in a data set from the Open Data Portal to their local R environment (e.g. R Studio). The write.socrata function, on the other hand, allows the user to upload an R data frame to a data set on the Open Data Portal. The user has multiple options for update functions, which include append, replace, and upsert. The user must provide a Socrata username and password before receiving access to read in a private data set from the Open Data Portal; credentials are not required to read in a public dataset. Once the user's R file contains all desired transformations and functions, a task scheduler such as Windows Task Scheduler can be used to automate the data set through regular updates.

Prerequisites

- Socrata account with the publisher or owner role
- Computer running a currently supported version of R
  - The latest version of R can be downloaded here.
- A currently supported R IDE, we recommend using RStudio
  - The latest version of RStudio can be downloaded here.

When is this tool a good choice?

Using the RSocrata package is a great way to apply your own personal transformations to a data set for upload to the Open Data Portal. As mentioned before, users will need to be familiar with the R programming language to get the most out of the RSocrata package. Luckily, there is lots of documentation on basic R capabilities available online. Again, like other issues with some tools in this guideline, to automate this tool, you must use a task scheduler. Windows Task Scheduler, unfortunately, will only run tasks when the computer is on, so the user must still be vigilant with some regard to keeping updates, or the script may be run on a server that is always on.

Process Overview

How to set up your R script to be able to use RSocrata functions:

1. Once you have RStudio open, click the File tab, and select New File > R Script. This will open a blank R script where you will write your code.
2. At the bottom of your screen in the Console, write install.packages("RSocrata").
3. Once the Console is done running installation, write "library(RSocrata)" at the top of your R script and run this line.

4. On the next two lines of your R script, write the following:

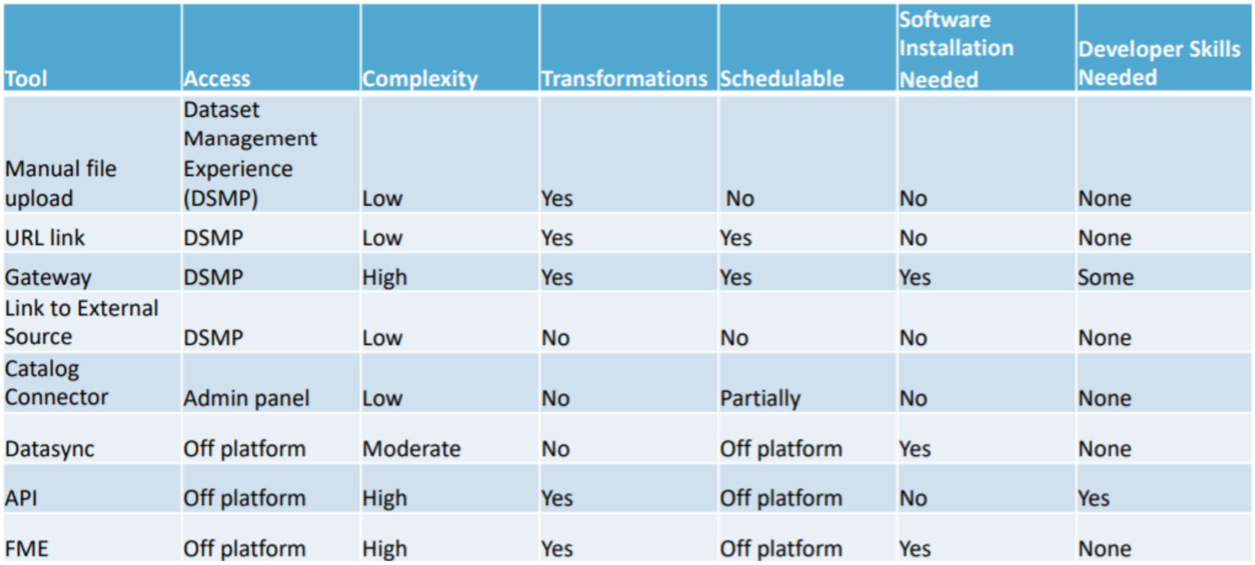
   ```
   socrata_username <- Sys.getenv("SOCRATA_EMAIL", "EMAIL_GOES_HERE")
   socrata_password <- Sys.getenv("SOCRATA_PASSWORD",
   "PASSWORD_GOES_HERE")
   ```

   Enter your Open Data Portal email and password where indicated above.
5. Run these two lines of code.
6. If there are no errors, then your R script is set up and ready to use all RSocrata functionalities!

For more information and examples on how to use the RSocrata package, see the documentation for the package here.

Automation Tool Comparison

The following two tables provide additional detail on some of the tools discussed in this document. The first compares the capabilities of various automation tools. The other provides some pros and cons for DataSync and Socrata Gateway. These tables are from a Socrata support document, which is available in the Socrata support article Overview of Ingres Methods.



Tool	Access	Complexity	Transformations	Schedulable	Software Installation Needed	Developer Skills Needed
Manual file upload	Dataset Management Experience (DSMP)	Low	Yes	No	No	None
URL link	DSMP	Low	Yes	Yes	No	None
Gateway	DSMP	High	Yes	Yes	Yes	Some
Link to External Source	DSMP	Low	No	No	No	None
Catalog Connector	Admin panel	Low	No	Partially	No	None
Datasync	Off platform	Moderate	No	Off platform	Yes	None
API	Off platform	High	Yes	Off platform	No	Yes
FME	Off platform	High	Yes	Off platform	Yes	None

Automation Tool Comparison

Method	Pros	Cons
Gateway	Once scheduling is set up, it will run automatically up to once a day. Supports a wide variety of data sources, including the US Census. Connections are created and managed through the Dataset Management Experience. Transforms, geocoding, and datatypes are easy to set in the Dataset Management Experience.	Users must install software on their own computer or server. The user environment (source system/machine/network factors) can affect the ease of setup. The user manages the connection. Can only replace data; appending data is not currently possible.
DataSync	Provides a basic user interface command tool. Used to import CSV files from a computer. Can import files over 4 GB. Can be used to replace all rows, append or upsert rows, or to delete rows. Used for "port jobs" - an easy way to copy Socrata datasets within or between domains. Can be run headlessly.	Free Socrata software, but it resides off the platform. Requires Java 8 (or newer) and installing DataSync, a small java application, on the user's machine. Must be scheduled by an external tool - Windows task manager, for example. No transforms are applied to data imported via DataSync