Open Data for Maryland State Agencies 101

Department of Budget and Management

Department of Information Technology

Governor's Office of Performance Improvement

Joshua Exler, Open Data Project Manager –
josh.exler@maryland.gov

v 1.2 – February 12 2016

Table of Contents

Introduction: For years, Maryland agencies have collected data and reported it to the Governor's Office, the Department of Budget and Management, and the Department of Information Technology. The data collected for these disparate needs often overlapped, but the timing of the collection and the methods of reporting differed. Moving forward, the plan is to consolidate data collection and reporting so performance can be reported in one place.

The goal is to make Maryland's Open Data Portal (http://data.maryland.gov) and MD iMAP, Maryland's Mapping Portal (http://imap.maryland.gov) the central locations to host State information.

Because http://data.maryland.gov is going to be the central data repository, and because the Governor's performance management programs will be using these data, adding and editing data to http://data.maryland.gov will be key.

There are three options for uploading datasets and maintaining them on http://data.maryland.gov:

- Direct Data Entry
- Uploading Excel Spreadsheets
- Connecting Source Databases

The current version of this document (v. 1.2) is being distributed as part of a hands-on tutorial for State employees on February 16, 2016. If you have any questions please contact josh.exler@maryland.gov.

● **Glossary:** Please refer to the table below for definitions of terms found in this document. If you have further questions, contact Joshua Exler: josh.exler@maryland.gov.

Data	The Open Data Act (State Government, Chapter 69, § 10-1401 through 10-1404) became effective June 1, 2014. It defines data as "…final versions of statistical or factual information that: (i) are **in alphanumeric or geospatial form** reflected in a list, table, graph, chart, map, or other non-narrative format that can be digitally transmitted or processed; (ii) are **regularly created or maintained** by or on behalf of a governmental entity; and (iii) **record a measurement, transaction, or determination or provide information** on government services, initiatives, and resources related to the mission of the covered governmental entity. (2) "Data" does not include draft versions of statistical or factual information that are used for internal analysis by a governmental entity.
Open Data	Open Data is any data collected by the State which can be provided to the public. According to the Open Data Act: " 'Open data' means data that, consistent with any applicable laws, rules, regulations, ordinances, resolutions, policies, or other restrictions including requirements or rights associated with the data**, a State entity:** **(i) has collected; and** **(ii) is permitted, required, or able to make available to the public**. (2) "Open data" includes contractual or other legal orders, restrictions, or requirements. (3) "Open data**" does not include data that if made public would**: (i) violate another law or regulation that prohibits the data from being made public; (ii) endanger the public health, safety, or welfare; (iii) hinder the operation of government, including criminal and civil investigations; (iv) impose an undue financial, operational, or administrative burden on a State entity; or (v) disclose proprietary or confidential information."
Open Data Portal	In these tutorials, "Open Data Portal" and "Portal" refer exclusively to http://data.maryland.gov, Maryland's statewide Open Data Portal. More

	broadly, the "Open Data Portal" is a product (Software as a Service, or SaaS), provided by the site's vendor, Socrata. In other contexts "Portal" can be a shorthand for MD iMAP, http://imap.maryland.gov, Maryland's mapping Portal.
Extract, Transform, and Load (ETL)	A three-step process to 1) Extract data from a source database, 2) Transform the data so that the format can be read by the destination database (in this case, the destination is a dataset on http://data.maryland.gov), and 3) Load the data to the destination database.
FME	Feature Manipulation Engine - a piece of software to write, schedule, and execute ETL functions. DoIT has two available licenses.
Data Series	A single measure or metric, usually plotted over time. Represented as a single column in a http://data.maryland.gov dataset. Called an "Attribute" in FME.
Record	For data measured over time, the number of records is equal to the number of observations (e.g., the number of months with data available). The number of records is the sample size (N). A record is represented as a single row in Socrata.
Metadata	Data about data. Literally translates as "next to data." This is information such as dataset name, publication date, attribution URL, dataset update schedule, and contact email address. A dedicated metadata page can be found for any http://data.maryland.gov dataset by appending "/about" to the end of its URL.

- **Pros and Cons of Uploading Data to http://data.maryland.gov: Three Methods.** Just as every state agency is unique, so is every data source. There is no one-size-fits-all solution for uploading datasets to data.maryland.gov. Broadly speaking, there are three options for uploading data to http://data.maryland.gov. The benefits and drawbacks of each are shown in the table below. For each new dataset you upload, consider which approach makes the most sense given your data, staffing levels, and technical capabilities.

Uploading to Data.Maryland.Gov – Three Methods		
Method Summary	**Pros**	**Cons**
Direct Data Entry	The easiest method to learn. It is as simple as typing data into a spreadsheet.	Data are manually inputted, then manually updated. In the long run, this method will require more work than will the other methods.
Uploading Excel Spreadsheets	Many existing agency reports are already regularly produced in spreadsheet format. Uploading a well-formatted Excel spreadsheet to http://data.maryland.gov is relatively straightforward.	Spreadsheets contain tabular data, but it is necessary to remove extra blank space, collapse three-dimensional data, and standardize data types to allow the files to be readable by http://data.maryland.gov. Updating an Excel spreadsheet and uploading it to http://data.maryland.gov whenever data are updated is a manual process.
Connecting Source Databases	Extract, Transform, and Load processes can be fully automated. DataSync can schedule basic CSV uploading jobs to http://data.maryland.gov. FME can convert between over 450 data formats, write to http://data.maryland.gov, write to http://data.imap.maryland.gov, and schedule automatic job execution and notifications.	Out of the three methods, requires the most up-front work. Custom built data connections by DoIT would be unique from agency to agency, and from database to database. Minimal code could be recycled from agency to agency.

- **Tutorial A: Direct Data Entry**

1. Sign in to your account at http://data.maryland.gov/login. If you do not already have an account please contact josh.exler@maryland.gov. Do not use the "Sign Up" link on http://data.maryland.gov, since this is intended for public users. Anyone can view data on http://data.maryland.gov, but only State users with accounts created by Josh Exler (DoIT) can create and edit datasets.

2. Once you log in you will be directed to your user profile page. Click the "Create a New Dataset" button.

3. Select "Design from Scratch." This will let you type data directly into a new dataset.

4. Fill out the metadata for your dataset in full. Please fill out as many metadata fields as possible.

▪ **Dataset Title** – The name of the dataset. Please give your dataset an accurate and precise title. Spell out abbreviations if needed.. Avoid "data" or "dataset" in the name and avoid all-caps. Make the title comprehensible to the general public. A good example is "Minority Business

Enterprise (MBE): Statewide Directory of Certified Businesses." A bad example is
"MBE_DATA_2016"

- **Brief Description –** Please include an overview of the dataset, data format, and source data. This is the only large text field in the metadata, so it also serves as a catch-all for other information about the dataset. This is the best place to spell out data collection, reporting, updating procedures, and workflows.
- **Category –** Select from the dropdown of available categories, e.g., Public Safety, Environment, and Health. If your dataset falls into more than one category, use your best judgement on which is most relevant.
- **Tags/Keywords –** The more tags you enter, the easier it is for site users to find your dataset. Please enter agency names and their abbreviations, as well as keywords about topics, subject areas, and programs.
- **License Type –** The majority of your agency's data is produced by state employees and should therefore be assigned a Public Domain license.
- **Data Provided By** – Please enter names of single or multiple agencies and organizations who produced the data. Please consider this field to be mandatory despite not being shown as 'mandatory' on this page.
- **Source Link –** URL to data itself or to an agency's website. Please consider this field to be mandatory despite not being shown as 'mandatory' on this page.
- **KML File, Shape File** – If you are working with geographic data, and either of these files are available elsewhere, please provide a URL here.

▪ **Place Keywords** – Please tag the dataset with any relevant geographic keywords. For statewide datasets this will usually just be 'Maryland.'

Create a new Dataset

Please describe your data.

* Dataset Title	Managing for Results - Maryland Department of Transportation (MDOT)
Brief Description	This dataset contains data from the Department of Budget and Management as part of the Managing for Results (MFR) program.
Category	Transportation
Tags / Keywords	transportation, maryland department of transportation, mdot, sha, state highway admini
	Enter one or more keywords separated by commas
Row Label	Year
	Describe what each row in the dataset represents (if applicable).

Licensing & Attribution

License Type	Public Domain
	Choose licensing terms for people who wish to use this dataset.
Data Provided By	Department of Budget and Management (DBM)
Source Link	http://dbm.maryland.gov/pages/default.aspx

Attachments

To upload attachments, first finish publishing this dataset. Then, click **ⓘ About** in the toolbar, and click Edit Metadata. You will then be able to upload attachments.

GIS Download

KML File	
Shape File	
* Place Keywords	Maryland

Time Period

* Update Frequency	Annually
Other Update Frequency - If frequency isn't included in list above, please describe it here.	
* Time Period of Content	Fiscal Year 2011 to Present
* Date Metadata Written	January 6 2016

Jurisdiction

* Jurisdiction	State of Maryland

Privacy & Security

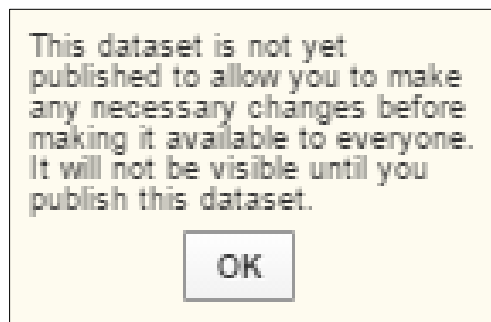|	
Privacy Settings	○ My data is **public**: it can only be modified by me, but everyone can view it ● My data is **private**
Contact Email	josh.exler@maryland.gov
	This address will not be displayed publicly. If left blank, it will default to your address

* Required Field

▪ **Update Frequency** – Please select from the dropdown list to indicate how often new data are released. Data tracked monthly would be considered monthly data, even if the data are made available at irregular intervals, e.g., 29 days after initial upload, then another update 32 days later, and so forth. Please do not select 'As Needed.' Please be accurate and precise. The text field below, "Other Update Frequency", is an option if you wish to describe for data update cycles in more detail.

▪ **Time Period of Content** – The dates over which data are tracked, e.g., "Fiscal Year 2011 to Present."

- **Date Metadata Written –** The most recent date you or any other user made changes to any of the metadata. Enter the current date for now. You can always update this in the future as you make changes to the metadata.
- **Jurisdiction –** If you are an employee of the State of Maryland, select 'State of Maryland.' The other options are for Maryland localities' Open Data microsites: Frederick Open Data, Mount Rainier Stat, and several in-development.
- **Privacy Settings –** Keep the dataset private for now. At a later step you will change this to public.
- **Contact Email –** Who will the public contact when they have questions about this dataset? Please consider this field to be mandatory despite not being shown as 'mandatory' on this page.

   Click "Next"…

5. You will see a spinning circle graphic for a few seconds. Then click the red "Finish" button.

6. You will be taken to the 'working copy' of your new dataset. A working copy is saved to the cloud storage of http://data.maryland.gov, but it is not yet published, meaning that it's neither public, searchable, nor available to any other users. The working copy is where you will enter and edit data. The site's vendor, Socrata, will warn you:

This dataset is not yet published to allow you to make any necessary changes before making it available to everyone. It will not be visible until you publish this dataset.

OK

Use the tools in the yellow 'Edit' column on the left to add the dataset's first data series, i.e., first column. Choose a data type. The most common options are Number, Percent, Money, Plain Text, and Formatted Text. All numeric data besides percentages and money should have the data type Number. For this example, the data series 'Year' will be data type Plain Text.

There are two main kinds of datasets on the Open Data Portal: time series datasets and individual records. Example of time series data include data reported monthly, quarterly, or annually. Examples of individual records include records for individual permits, traffic accidents, or persons.

7. For this tutorial we will enter data for a few annually reported data series. The sample datasets in these tutorials are taken from agencies' Managing for Results data.

Double click in the first row in the 'Year' column to begin typing there. To finish data entry in that cell, hit Enter.

8. Double click in the cell below, to enter the next year, "2012 Act." Double clicking in the cell creates the next record.

9. Continue to enter additional years into the 'Year' column.

10. Repeat steps 8-11 to create a second data series, i.e., a second column. The first Department of Transportation metric from its Managing for Results data is 'Vehicle miles traveled on MDTA facilities (in millions)', which should have the data type 'Number.'

11. Watch out! For the 'Number' data type, non-numeric characters – like commas – are rejected. Any time your data type is incorrect, the cell will show up as light gray and italicized. For example, including commas in the data entry for 2011 Act. and 2012 Act. produces the following:

← If you see data like this (italicized and light gray), it means there is a problem with your data's formatting, such as unexpected commas.

12. After you have added as many columns as you need, and entered all of your data, click "Publish Dataset." This will save your Working Copy as a new published dataset. Your new dataset will live at a static URL which you can bookmark if you wish.

13. When you are ready for your dataset to be public and searchable, use the Privacy tools in the brown 'Manage' menu.
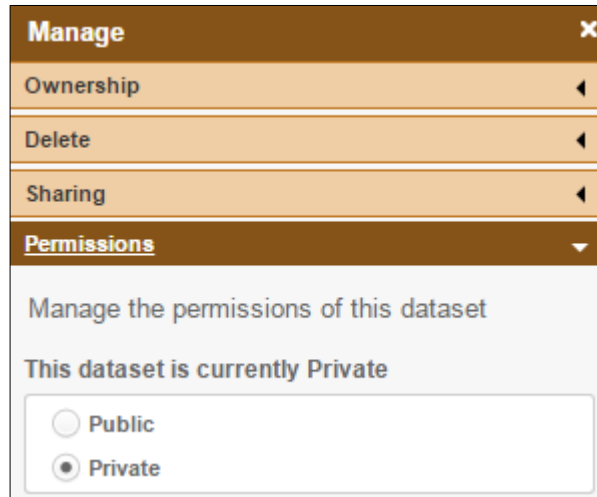
When you want to make changes to your dataset, such as adding new data or a new data series, click the "Edit Dataset" button in the yellow 'Edit' menu to create a new working copy. Each new Working Copy will have a new URL generated, but every time you publish your dataset, the published version stays at its permanent location.

- **Tutorial B: Uploading Excel Spreadsheets.** Instead of performing data entry by hand, you can upload Excel spreadsheets or CSV files to new datasets on http://data.maryland.gov. You can also use Excel to update an existing dataset. This tutorial covers uploading a new dataset using an existing Excel sheet. Some of these reference steps from Tutorial A, instead of repeating the same information twice.

1. Identify an Excel spreadsheet to upload. The hard part of this approach is getting your spreadsheet to match a data format which http://data.maryland.gov is expecting. This is the 'Transform' step in the Extract, Transform, and Load (ETL) process. A http://data.maryland.gov dataset works best with tabular data, i.e., flat data which looks like a data table. More complicated data structures, e.g., XML files with multiple layers of nested attributes, will need to be flattened before upload. Please contact DoIT and josh.exler@maryland.gov if you are ambitious and want to upload datasets with relatively complicated data structures.

2. Work in Excel to make sure the dataset's format is readable by http://data.maryland.gov. Please follow the following guidelines:

- The first row should contain the names of each data series.

- Each data series is represented as a column.
- Each record, e.g., each month of data, is represented as a row. It is not possible to switch this, i.e., you cannot use rows as data series nor columns as records.
- Each data series has its data all in the correct format. e.g., a data series with the data type of 'Number' should not have any text in it – no asterisks, no explanatory notes, etc.
- Please delete any other cells in the spreadsheet besides those in the data series. Erroneous cells will confuse http://data.maryland.gov upon dataset upload.
- Save the spreadsheet as a CSV. This file type is smaller than is XLS or XLSX, and does not contain any styling information. CSV is a more reliable file type for uploading to http://data.maryland.gov.

While it may look like a lot of formatting work, all you are doing is making your Excel spreadsheet look like a simple data table, and look like the format you are going to see on http://data.maryland.gov. Here is an example, showing part of an example spreadsheet (with Maryland Department of the Environment MFR data):

	A	B	C	D
1	Year	A responses issued vs processed within app	onmental emergenci	
2	2011 Act.	62.%		
3	2012 Act.	75.%		
4	2013 Act.	65.%	82.1%	571
5	2014 Act.	80.%	90.6%	537
6	2015 Act.	85.%	92.1%	492
7	2016 Est.	85.%	90.%	500
8	2017 Est.	85.%	90.%	500
9				

3. Log in to your http://data.maryland.gov account (see steps 1 in Tutorial A for more details).

4. On your user profile page, click "Create a New Dataset."

5. Select "Import a Data File." This will let you upload your CSV file.
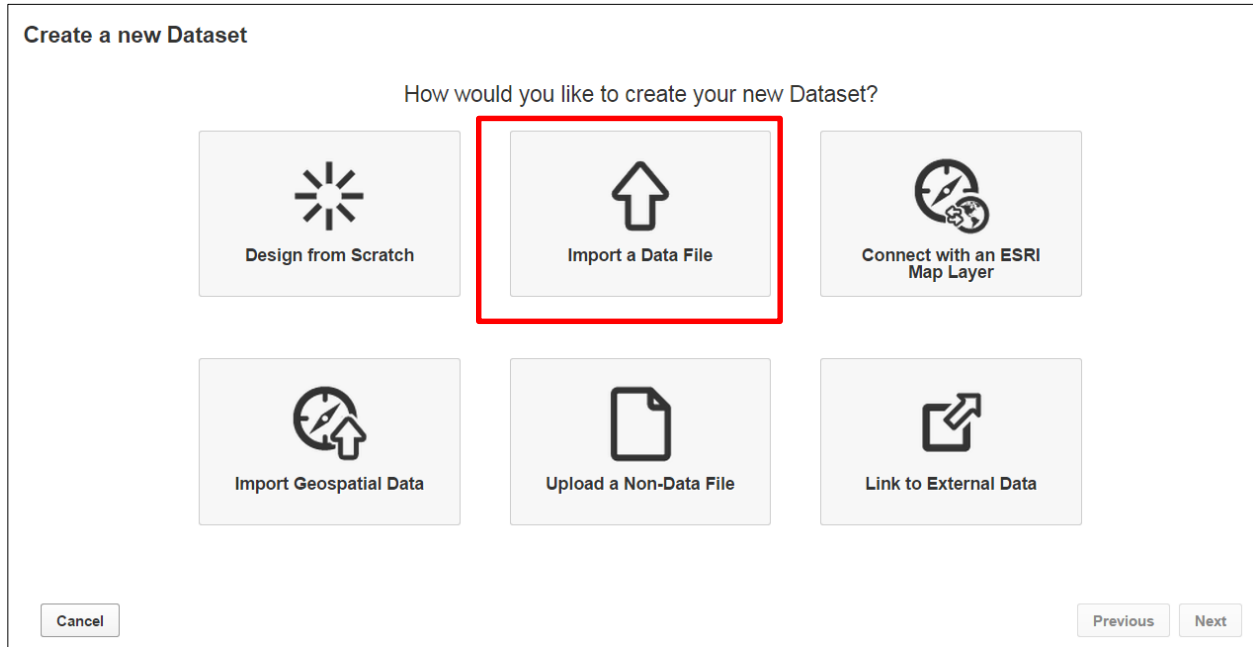
6. You can choose a file on your local disk, or a file already online at another location. Once you have selected your file, click "Next."

7. Choose data types for each data series you are importing. The most common options are Number, Percent, Money, Plain Text, and Formatted Text. All numeric data besides percentages and money should have the data type Number. The data series 'Year' will be data type Plain Text. The bottom of this screen will show you a preview of the resulting http://data.maryland.gov dataset.

8. If you have any warnings or errors at the bottom of the screen, try to work through each. The most common problems will come from data type mismatches. Click "Next" to begin your data upload.

9. While your data is uploading, you will see a spinning circle graphic. Once your data are imported. Click the red "Finish" button.

10. You will be taken to the 'working copy' of your new dataset. A working copy is saved to the cloud storage of http://data.maryland.gov, but it is not yet published, meaning that it's neither public, searchable, nor available to any other users. The working copy is where you will enter and edit data. The site's vendor, Socrata, will warn you:

11. Follow steps 12 and 13 from Tutorial A to publish your Working Copy, then to make your dataset public.

This tutorial has covered how to create a new dataset by uploading an Excel spreadsheet. To update an existing dataset using data from an Excel sheet, first, click "Edit Dataset" to make a new working copy. Then, once your new working copy has loaded, use Edit > Append & Replace. This will bring up the dataset "Append and Replace" wizard, which will walk you through the steps of either appending data to your dataset from a spreadsheet (i.e., adding new rows onto the bottom of the dataset), or replacing your dataset's data from a spreadsheet (it is expected that the data format of the spreadsheet will match the format of the http://data.maryland.gov dataset, meaning that the column names and data types should be the same).

- **Tutorial C: Connecting Source Databases.** Compared to the methods detailed in Tutorials A and B, connecting data directly from a source database does not have step-by-step instructions. It requires relatively more technical expertise and relatively more assistance from the Department of Information Technology (DoIT). There are two main methods available for connecting source databases, and overviews of both are included below: Using Socrata DataSync to schedule data uploads to the Open Data Portal – and using Feature Manipulation Engine (FME) to create and schedule Extract, Transform, and Load (ETL) functions. These two applications have been briefly mentioned earlier in this document, but this tutorial presents additional details on the capabilities, advantages, and limitations of both.

- *Using Socrata DataSync to Schedule Data Uploads.* If you have access to source data in tabular formats – especially in CSV format – Socrata, the vendor of the Open Data Portal, offers a lightweight Java application to author and schedule data upload jobs. Socrata provides the following description for DataSync on the DataSync documentation site: *"DataSync is an executable Java application which serves as a general solution to automate publishing data on the Socrata platform. It can be used through a easy-to-use graphical interface or as a command-line tool ('headless mode'). … DataSync takes a CSV or TSV file on a local machine or networked hard drive and publishes it to a Socrata dataset so that the Socrata dataset stays up-to-date. DataSync jobs can be integrated into an ETL process, scheduled using a tool such as the Windows Task Scheduler or Cron, or used to perform updates or deletes in batches. DataSync works on any platform that runs Java version 7 (a.k.a version 1.7) or higher (i.e. Windows, Mac, and Linux."*[1]

  Socrata's **Quick Start Guide** for DataSync contains detailed, step by step instructions for installing the application, and scheduling a job. It is important to note that this process requires that you create an Application Token, which is necessary for the application to be able to make REST-ful queries behind the scenes, using the Socrata Open Data API (SODA).

  DataSync can perform several basic data transformations by setting options to, for example, remove white space, or change character encoding. Options can be setting by setting up a Control File. However, for data where cleanup and significant data transformations are required, these changes be made directly in your CSV file, before setting up your ETL function in DataSync.[2]

---

[1] "Getting Started with Socrata DataSync." https://socrata.github.io/datasync/

[2] An ETL function in DataSync is called a "Socrata Integration Job (SIJ)." SIJ jobs are saved in .sij files using DataSync. DataSync is the only application which can open .sij files.

If your agency's workflow involves using Microsoft Excel to generate XLS, XLSX, or CSV files for upload to the Open Data Portal, a hybrid solution using both Excel and DataSync is often the most efficient option. Once you have your data in a tabular format in an Excel spreadsheet, save the file as a CSV then use DataSync to automate the uploading of the CSV to the Open Data Portal. DataSync can essentially replace and automate the data "Append & Replace" tool detailed in Tutorial B.
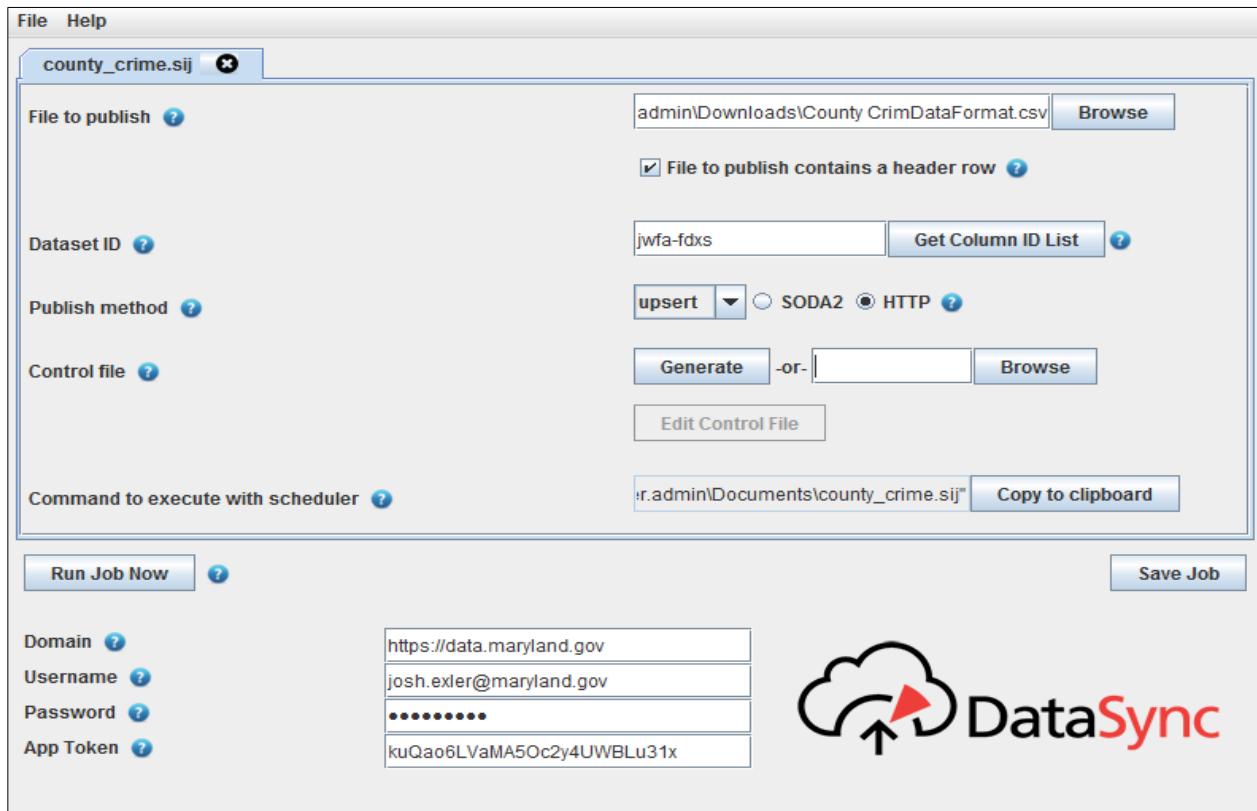
A screenshot from DataSync v. 1.5.4.

- *Using Feature Manipulation Engine (FME)*. Feature Manipulation Engine is a piece of software to write, schedule, and execute ETL functions -- allowing almost any data source to be automatically reformatted and automatically uploaded to the Open Data Portal. DoIT has two available licenses. FME can convert between almost any data format, about 350 in total, and can script and automate virtually any type of data transformation. FME has over 475 built in "transformers", which act as steps in the ETL process. A typical FME workspace involves reading data from an internal or external location using a "reader," e.g., from a spreadsheet or database. Data is then transformed using any number of transformers, including the ability to add custom transformers using Python. Finally, data are loaded to either a local or remote location. Because writers for Socrata and ESRI products are included in FME, the software can script and

schedule ETL jobs both for Open Data Portal datasets and MD iMAP datasets. ETL jobs can be scheduled to automatically execute at set intervals, and automatically notify set email addresses when ETL jobs succeed and fail.

For all ETL jobs which are scripted and automated using FME, DoIT plays a role. The first step for any agency interested in automatic the ETL process for their data uploads to the Open Data Portal using FME is to contact DoIT at service.desk@maryland.gov. You should have source data available before contacting DoIT to begin this process. Unlike with DataSync, however, you do not need to worry about the file type nor the data format, since FME can import data with almost any extension and format.

Much like DataSync, FME uses the Socrata Open Data API (SODA) behind the scenes. FME generates Application Tokens automatically as part of the execution of jobs which write to the Open Data Portal, so you do not need to worry about manually generating an Application Token.

DoIT is able to load data sources into FME if your agency is able to make your data available to DoIT through FTP, or uploaded to the Google Drive folder associated with your State of Maryland Google Apps account, or even emailed to DoIT. DoIT can provide advice on how to export data from your agency's source databases automatically, but it is usually infeasible for FME, running on a DoIT server, to pull data directly from another agency's network, without considerable additional difficulty, e.g., setting up your agency's firewall to allow access to FME.

As an example of how this process can work: the State Department of Assessments and Taxation (SDAT) provides DoIT with access to data from its C.A.M.A. (Computer Aided Mass Appraisal) database by dropping updated .zip files in a folder to which DoIT has FTP access, with new data provided every several days. Since FME can connect to FTP, these files are automatically downloaded and used in an FME workspace, resulting in significant data transformation and automatic uploads to the Open Data Portal. ETL jobs which are saved with FME to a DoIT network drive are then scheduled to automatically execute monthly, all using tools available in FME.

This document is continually updated in response to feedback received by State employees. If you are unclear about the steps in any of these tutorials, please contact Joshua Exler at josh.exler@maryland.gov for clarification, and so that future versions of this document can be improved. Once you have uploaded your data to the Open Data Portal and you are comfortable updating your data using the methods in this document, your dataset will be reviewed by the administrative team at DoIT and made public. To start playing around with your data, making graphs, maps, and filtered views, first review these introductory training videos from Socrata at https://opendata.socrata.com/videos