# Background:

The following tools were built in ArcGIS 10.5.1 and Microsoft Excel 2016.

This document was created by GISinc August of 2017. The latest update was made 6/25/2018 by Jason Ford.

The process was developed to check state-supplied address data for basic quality control, then transform it and load it into a standardized national address database format. The supplied data should be as close as possible to the same schema as the national address database, and therefore should only use minor translations and attribute mapping. Values that are not mapped will be pushed to a flagged feature dataset for review and repair in hopes that over time the data will improve across all levels. The document outlines the procedures for setting up and processing the submitted data. For an overview of the ETL flowchart, see the appendix at the end of this document.

# Setup:

Enable the following extensions: Background geoprocessing, Data Interoperability

Open ArcMap or ArcCatalog and connect to the folder containing the Python Toolbox  
 (…#USDOT\\_Scripts\USDOT\_NAD.pyt)

Expand the toolbox (it may take a moment for the tools in the toolbox to show up)

# Process a New State Submission

## Setup Folders for New State Submission

Within the head folder “#USDOT\”, add the following folders and include the submitted geodatabase:

🗁 ST (State two letter code)

🗁 ST\_YYMMDD (date of submission)

🗁 PreETL\_QC

🛢 Submitted\_geodatabase.gdb

∴ Source\_Address\_Pts

🗁 Result\_QC

Copy the empty NAD\_Template\_[date].gdb (from #USDOT\Nad\_Template\) to the [ST]\_YYMMDD\Results\_QC\ folder location, and rename it to [ST]\_ETL\_Results.gdb. Rename the empty feature class to [ST]\_postETL.

🗁 Result\_QC

🛢 ST\_ETL\_results.gdb

∴ ST\_postETL

## Run the Feature Count tool:

Input: Point the tool to the feature class to be evaluated (i.e. Submitted\_geodatabase.gdb\Source\_Address\_Pts)

Output: Provide an output location and filename ending in .csv (i.e. ST\ST\_YYMMDD\PreETL\_QC\ST\_FeatCount.csv)

Open ..\#USDOT\\_NAD\_template\**Master\_Domain\_frq.xlsm** and enable macros.

On the Dashboard tab, press the button to select the CSV (ST\_FeatCount.csv) that you generated from the Feature Count tool. Then press the button to run the macro.

Open & format CSV the CSV file:

Select the CSV file:

Select the CSV file:

After the file opens, you’ll need to add an asterisk (\*) to the end of the field name for each required field. The list of required fields is listed below.

To quickly append ‘\*’ to a value, follow these steps: Select the cell with the field name or a required field and press [F2] then [\*] to add an asterisk to the end of the string and press [Enter].

This will **bold** the required fields, and  **highlight**  the % and count if the required field was not completely populated.

10 required fields in the NAD schema:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State\* | StreetName\* | Longitude\* | NatGrid\_Coord\* | Source\* |
| County\* | Add\_Number\* | Latitude\* | GUID\* | LastUpdate\* |

## Run Domain Checks using the Frequency Tool

### Processing Data that ***is*** already in the NAD Schema:

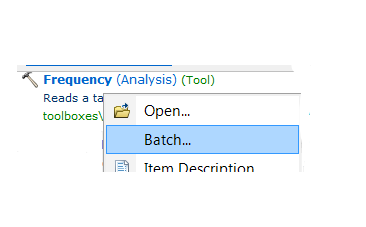
Run the **Frequency** Tool (within the \_Scripts\Toolbox) using the same input

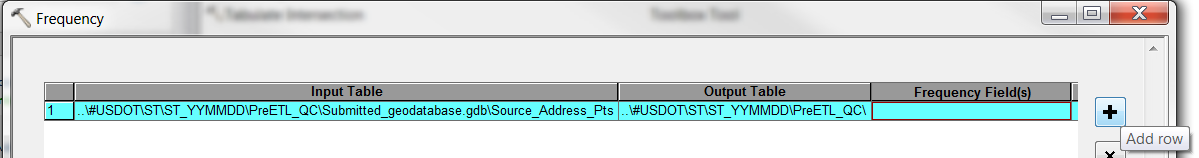
Input: Point the tool to the feature class to be evaluated (i.e. Submitted\_geodatabase.gdb\Source\_Address\_Pts)

Output: Provide the corresponding output folder for that state submission (i.e. ST\ST\_YYMMDD\PreETL\_QC\)

### Processing Data ***not*** provided in NAD Schema:

Copy the file path for the folder destination (i.e. ..\#USDOT\ST\YYMMDD\PreETL\_QC\)

Run the Frequency Tool from the ArcCatalog toolbox in Batch mode (right-click the toolbox tool and select Batch)

For the first line item, fill in the Input Table (i.e. Submitted\_geodatabase.gdb\**Source\_Address\_Pts**).

Then replace the destination path with the path in your clipboard (i.e. ..\#USDOT\ST\YYMMDD\**PreETL\_QC\**).

Now, while the line item is highlighted, use the plus sign on the right to duplicate the line. This will keep the same input and destination folder value for all line items.

Next, select the destination path again and add the name of the output file starting with a prefix of “frq\_” followed by the corresponding feature class in the NAD schema and ending with “.dbf” extension (ex: Output Table ..\#USDOT\ST\YYMMDD\PreETL\_QC\**frq\_StN\_PreDir.dbf** ).

Finally, set the corresponding frequency field from the source database. For example, the StN\_PreDir (Pre-Direction) field in the NAD schema may correspond with the pre-direction field in the source called PD.

### Process the Results

Now open the Master\_Domain\_frq.xlsm and enable macros if it is not already open.  
Location: ..\#USDOT\\_NAD\_template\**Master\_Domain\_frq.xlsm**

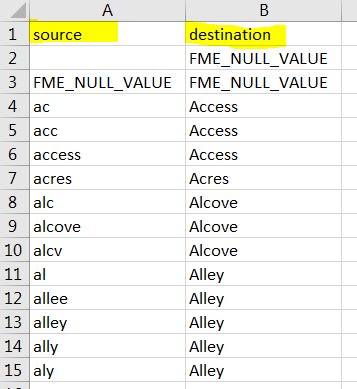
On the Dashboard tab, press the [Select a folder:] button to select the folder containing the dBase files generated from the Frequency Tool. Then press the [Merge & format dBase results] button to run the macro.

This will open each dBase file and compile the contents in a single spreadsheet. The data values will then be compared to the values in the corresponding tabs in the Master\_Domain\_Frq.xlsm workbook. If the value is not found, the cells will  **highlight**  red. Note: <null> or empty values in the dBase files are indistinguishable and may need to be verified in the source database.

Save the file with a name similar to “ST\_DomainChecks.xlsx” within the …\ST\YYMMDD\PreETL\_QC\ folder so you can reference back to it if needed in the future.

Determine if any values in Red should be added to the ‘repairable’ values in the csv reference tables used in the ETL to map source values to the corresponding destination domain value. If there are values that are not in the Master list but are verified to be accurate values, they can be added to the Master\_Domain\_Frq tab for the corresponding field, added to the csv mapping tables, added to the domain CSVs, and subsequently loaded as new domain values in the NAD\_template.gdb

# Make Updates to Domains:

Update the mapper\_[Field Name].csv (ensure source values are in lower case and destination values are in proper case). This file contains known abbreviations and misspellings and their matching domain values in the NAD schema.

Location: …\#USDOT\\_Scripts\Mapper\_[Field Name].csv

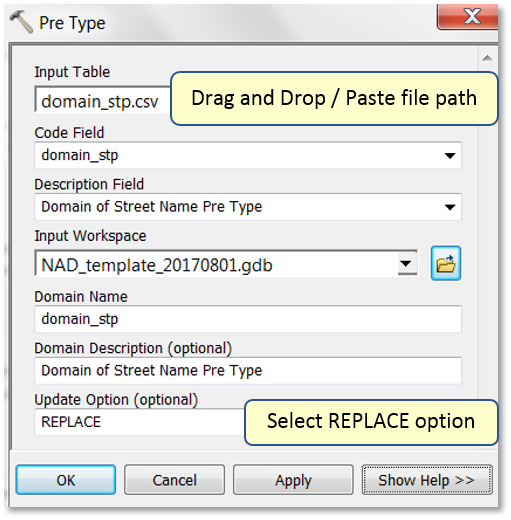
Update the Master\_Domain\_frq.xlsm (ensure source values are in proper case/title case). If you update a prefix type, don’t forget to also update the post type as well. When processing the results of the frequency and domain check tools, this file references the values listed in each domain tab to highlight values it cannot match. The lists on these tabs do not contain abbreviations or misspelled domain values.

Location: …\#USDOT\\_NAD\_template\Master\_Domain\_frq.xlsm

Update the domain\_[Field Name].csv (ensure source values are in proper case/title case). This should match the Master\_Domain\_frq.xlsm list for the corresponding tab. These are the CSV files that are used to update the domain tables in the file geodatabase schema.

Location: …\#USDOT\\_NAD\_template\domain\_[Field Name].csv

Run Table to Domain tool using the NAD\_Template.gdb and the domain\_[Field Name].csv



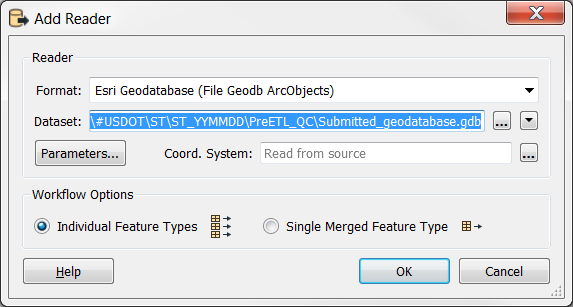
# ETL Process

Locate the Spatial ETL Tool. Right-click and select edit.  
(🗁 \_Scripts / 🗂 Toolbox.tbx / 🛠 Spatial ETL Tool)

Locate the ETL.fmw file (…\#USDOT\\_Scripts\ETL.fmw) and copy it to the state submission folder, then rename the file with the State two letter code. Next drag the file into the main tab of the FMW workbench to load the file.  
(🗁 ST (State two letter code) / 🗁 ST\_YYMMDD (date of submission) / ST.fmw)

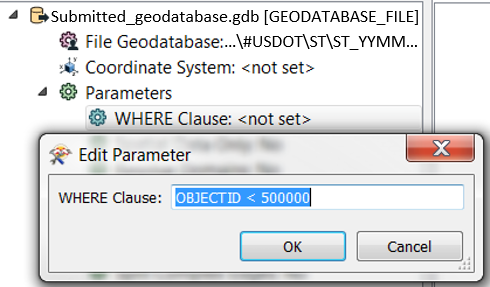
## Setup for processing a new state:

Add reader:

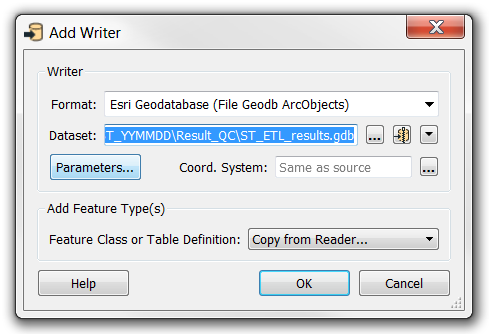


I have found the best success with filtering for ~500,000 records at a time through the ETL.

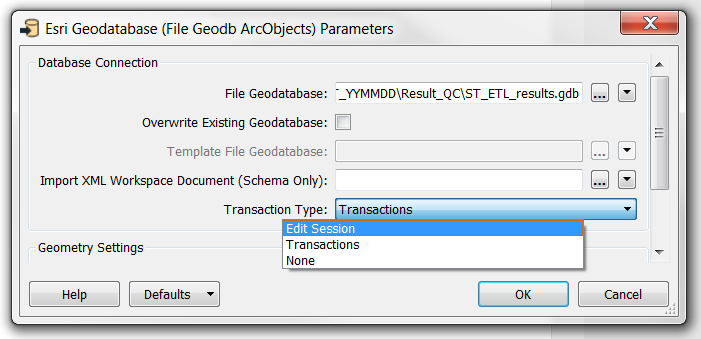
Edit the WHERE clause: **OBJECTID < 500000**

****

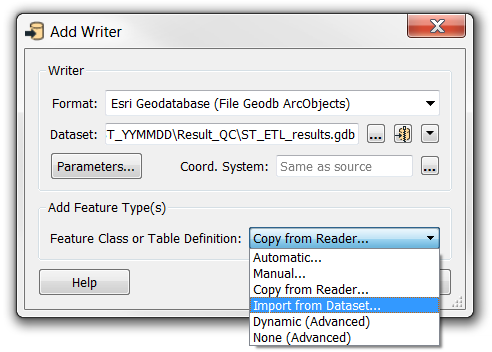
Add writer:



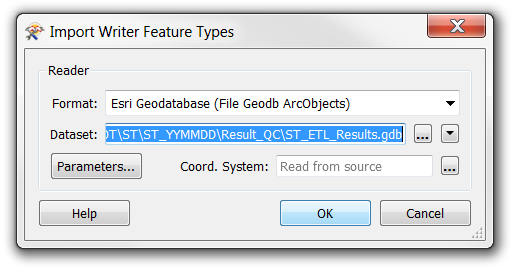
Then click Parameters and under the Database Connection, set the Transaction Type to Edit Session.



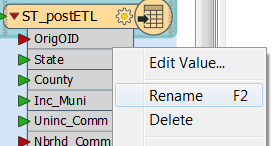
Under Add Feature Types, select Import from Dataset… from the drop—down option. Click OK.



After it begins to create the writer, a new prompt will appear to Import Writer Feature Types. By default, the dataset should still be set to match the Destination Dataset you just selected. (Since this was created from the NAD template, this ensures the feature type will use the NAD schema. Click OK.

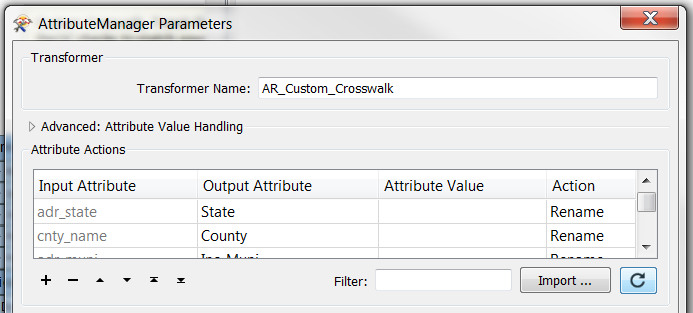


Rename the attribute in the writer from OBJECTID to OrigOID

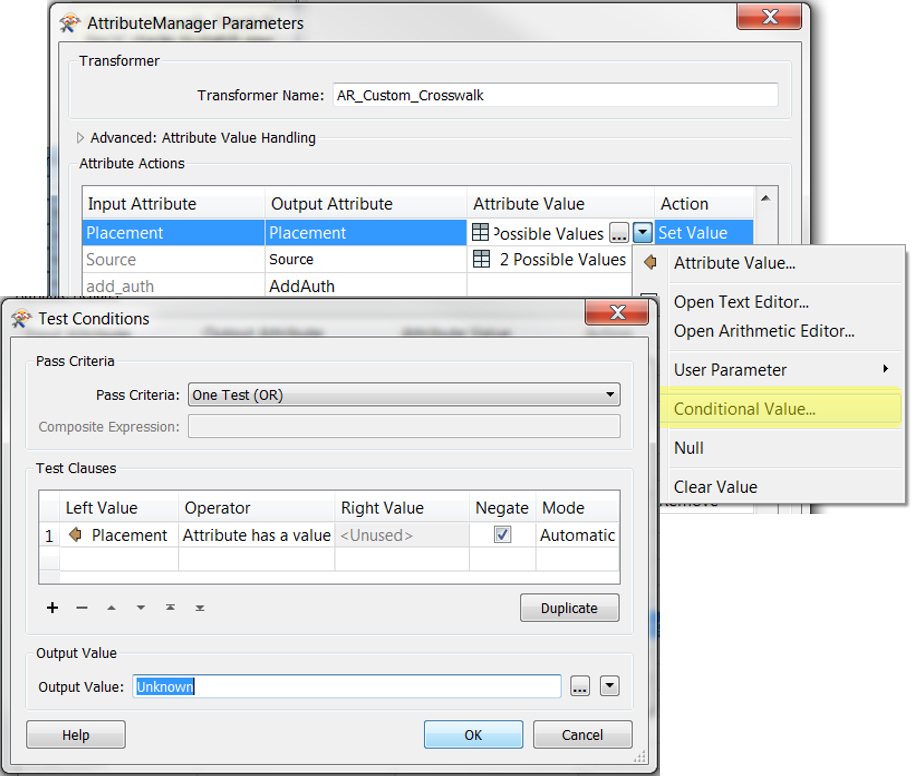


Edit the [ST]\_Crosswalk (Attribute Manager) transformer:

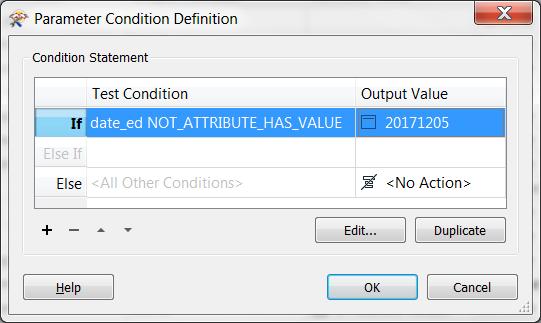
You can click the refresh button in the lower right to import the incoming schema to the transformer. From there you can identify and set the corresponding fields for Input Attribute and Output Attribute.



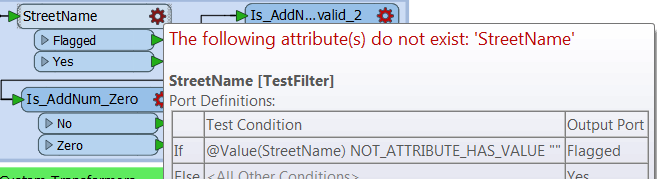
There are a few attributes that can be handy to add a conditional value if the attribute value is empty/null such as Addr\_Type, Placement, Source, LastUpdate. Example: If Placement has no value, set value ‘Unknown’



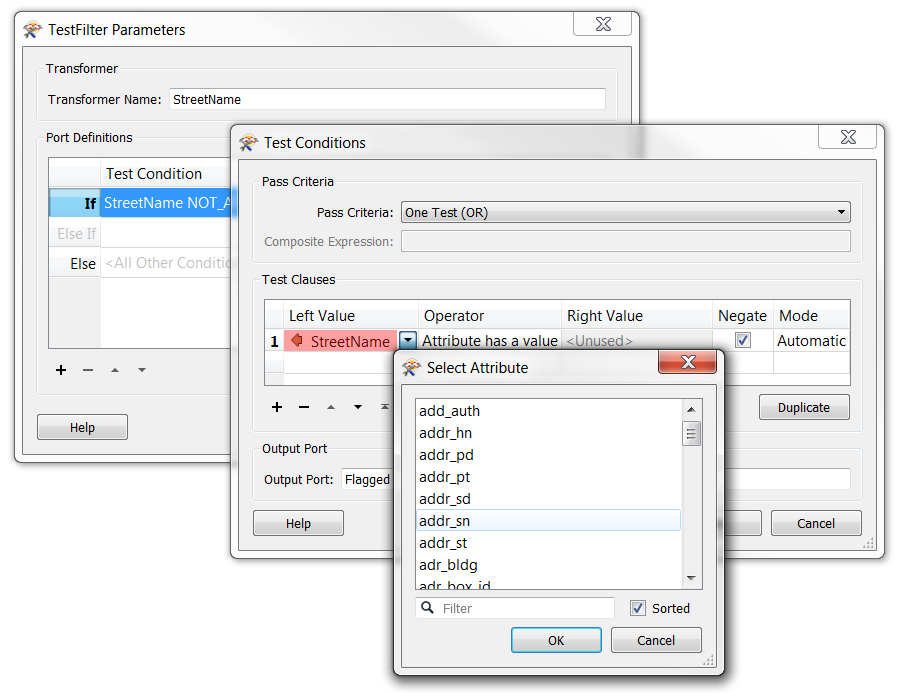
When updating the conditional value for LastUpdate attribute, use the date format YYYYMMDD.



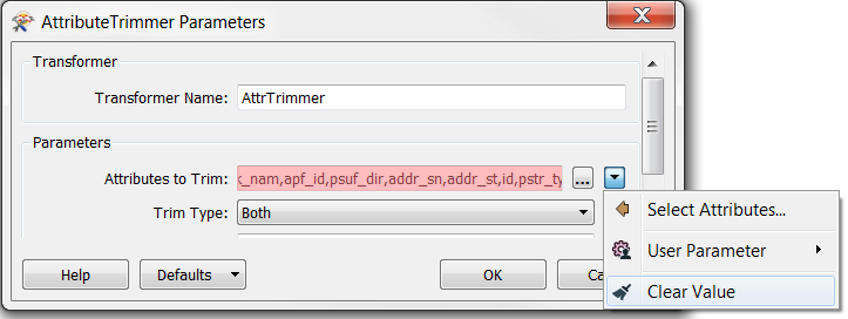
Repair any broken transformers (the gear on the right side of the transformer will be red) that may be a result of the source data not containing the same field names as the NAD schema. Common transformers that need to be updated when using state-specific source schemas are: Duplicate Filter, StreetName, Address\_Number testers, etc.



Double-click the broken transformer and replace parameters with correct corresponding attribute fields from the source geodatabase.



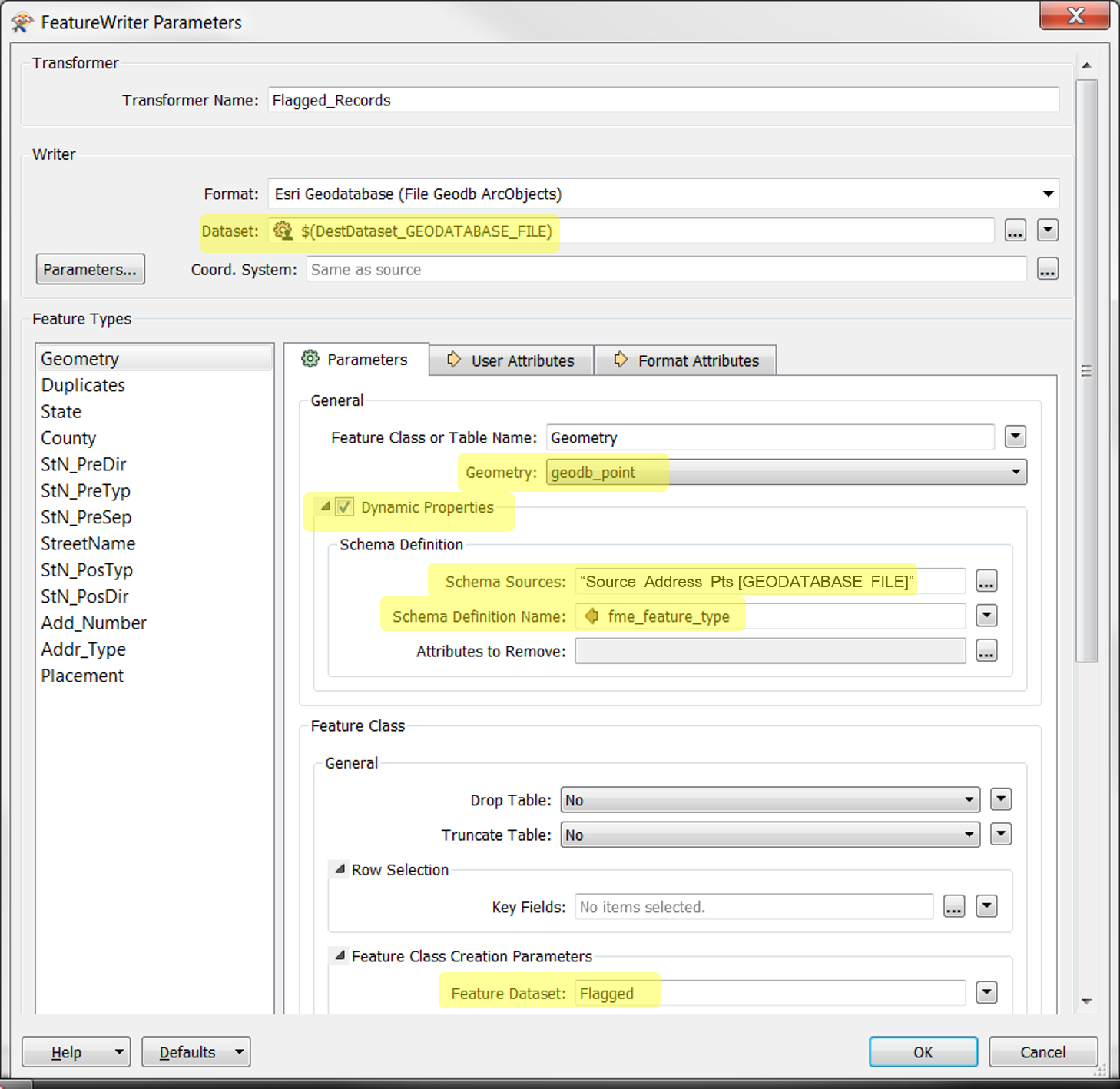
Other transformers such as the Attribute Trimmer are applied to all fields, so they can usually be repaired by first clearing the *Parameters: Attributes to Trim* and selecting OK, then re-open the parameter and within the Select Attributes menu, check the ‘select all’ box. Click OK and the transformer should no longer have a red gear.

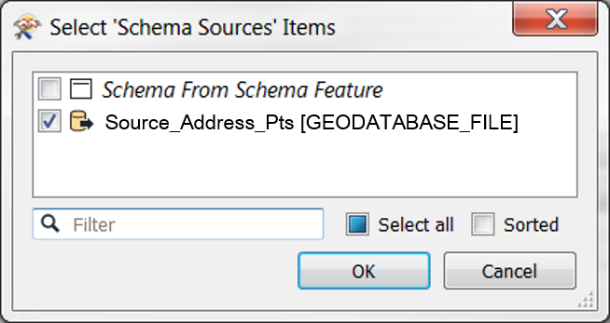


Connect any state specific transformers such as GUID or the set of transformers for NatGrid\_Coord, Latitude and Longitude (since they require reprojection first). You can check the results from the PreETL\_QC checks for feature count and domain checks to decide what needs to be repaired or calculated for the source dataset during the ETL.

In the Flagged\_Records transformer (FeatureWriter), check all feature classes on the left side and verify they all have the following settings:

* Ensure *Writer: Dataset:* **matches the main dataset writer** you just added. (We want to be sure the flagged *and* accepted records all get written to the *same* geodatabase)
* *Parameters: General:* **Dynamic Properties is Checked** (if this was unchecked, it will likely reset the featureclass name but you can select the drop-down arrow and clear value to set it back)
* Ensure *General: Geometry:* **geodb\_point**.
* Then, in the *Dynamic properties: Schema definition* section, set the schema sources to **only match the source reader** (uncheck *<Schema From Schema Feature>*),
* and set the *Schema Definition Name:* **fme\_feature\_type**
* Set *Feature Class Creation Parameters Feature Dataset:* **Flagged**





Now that the Readers, Writers, State-specific Transformers, Crosswalk, and Broken transformers have been set up and repaired, It’s time to run the ETL. I like to verify a few things before running the ETL.

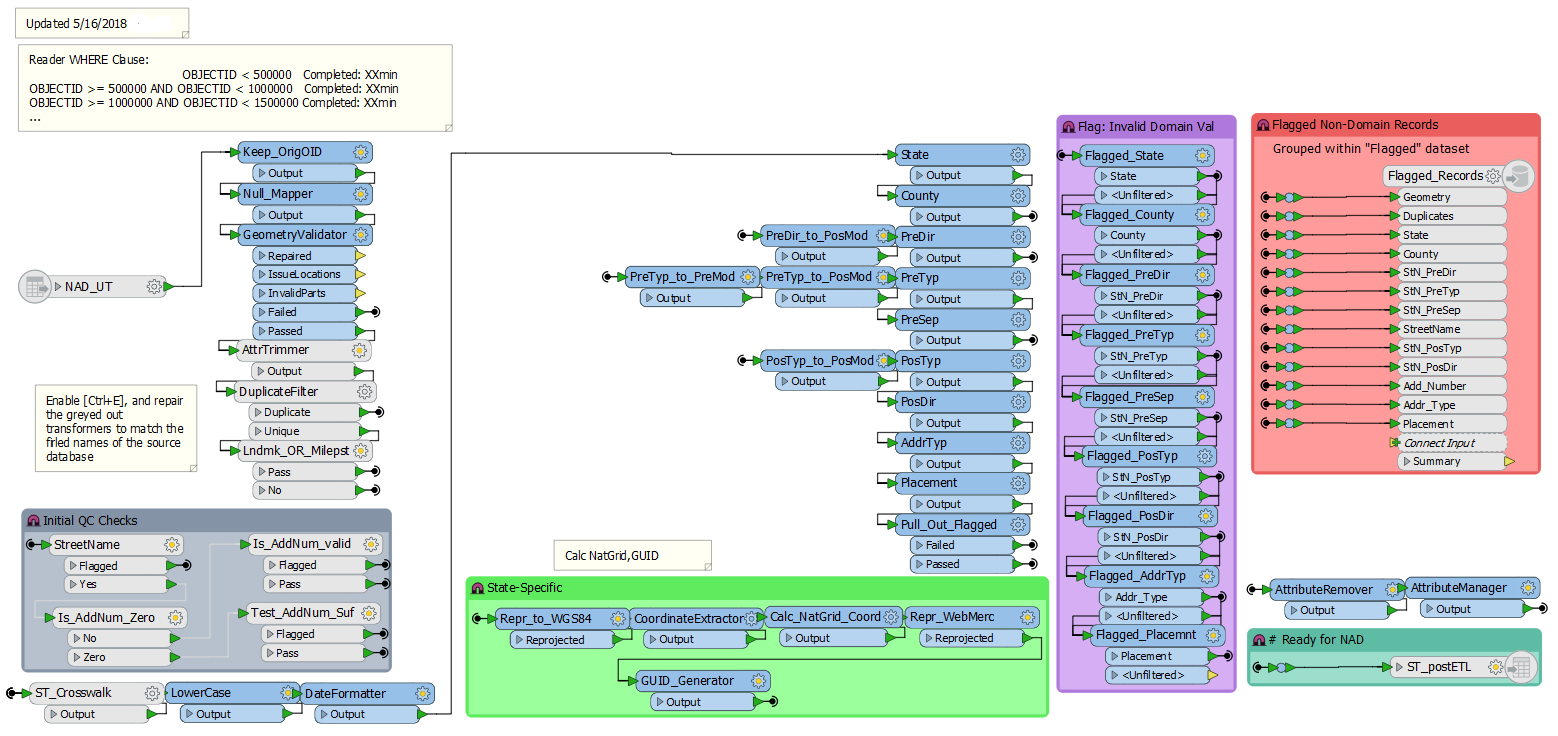
-Reader: WHERE clause: is set to filter for 500k records (Side Note: This can affect locating duplicates from one batch of 500k records to the next)

-Writer: Transaction Type = Edit Session

Then click Run.

# Appendix:

## The ETL generally does the following processes:



* Rename the OBJECTID to OrigOID. This is done to make it easier to join the output back to the source database at the state level. The OrigOID field does not exist in the NAD schema, so thie field will be dropped when loading from the output of the ETL to the NAD database.
* Map empty/missing/[space] values to Null
* Validate geometry
* Trim / Cleanup leading and trailing spaces from all attributes
* Duplicate Record Checks:
  + Longitude
  + Latitude
  + AddNum\_Pre
  + Add\_Number
  + AddNum\_Suf
  + StN\_PreMod
  + StN\_PreDir
  + StN\_PreTyp
  + StN\_PreSep
  + StreetName
  + StN\_PosTyp
  + StN\_PosDir
  + StN\_PosMod
  + LandmkPart
  + LandmkName
  + Building
  + Floor
  + Unit
  + Room
  + Addtl\_Loc
  + Milepost
* Is one field required? Is there a prerequisite? (Ex: address number and street name is required unless the record is a landmark or milepost record)
* Crosswalk fields from the source to the NAD schema, and supply the text value for the Source field based on the state that is being processed.
* Convert all domain fields to lowercase values. This could be done as upper or lower, but the ETL has been built to use lowercase in the lookup tables, be consistent if you decide to change this. Because there are common abbreviations that are supplied as a mix of upper, lower, and mixed case, converting all values to a single case, then using single-case source lookup values allows for the lists to be much shorter and easier to maintain.
* Format date fields, this is applied to attempt to catch any data fields that may have been supplied in a different format.
* Domain checks (known abbreviations, mixed-case, etc will be mapped to match the NAD domain values). These transformers are static and do not update when the lookup table CSV files are updated. The advantage of the AttributeValueMapper in this case is that it allows output to a new field (essentially just prefixed with an underscore) and will over-write any unmatched values with the value of your choice; in this case, “FLAGGED.”
* At this point, there will be two similar fields for each domain. The original domain field containing the original source value, and a new field prefixed with an underscore containing the expanded/corrected domain value OR the value “FLAGGED.” The FLAGGED records are then parsed out and passed through a series of filters to save out the original domain field values to the appropriate error type within the output Flagged feature data set. The valid values in the “\_[domain]” fields will continue through the rest of the ETL process.
* Now that all flagged values have been removed, any state specific transformers and transformers such as calculating the NatGrid\_Coord which requires reprojection can take place.
* Finally, the invalid original domain fields are removed, and the “\_[domain]” fields are renamed to remove the underscore prefix.
* The accepted records are written to the [ST]\_postETL feature class.

## Folder Structure

🗁 \_Documentation

🗖 Presentation

🖺 Report\_Template.docx

🗁 \_NAD

🛢 NAD.gdb

∴ NAD

🖹 Contacts.txt

🖹 Source.txt

🗁 \_NAD\_template

🛢 NAD\_Template.gdb

∴ NAD

🖹 domain\_county.csv

🖹 domain\_placement.csv

🖹 domain\_...csv

🗠 Fields.xlsx (List of attribute fields and formats)

🖹 NAD\_template\_170801.XML (Aug 1, 2017 NAD Template XML Schema)

🗠 Master\_Domain\_Frq.xlsm

🗁 \_Scripts

🛢 Reference GDB

❖ Counties

❖ Counties\_WebMerc\_3857

🗂 Toolbox.tbx

🗗 Calc\_County

🗗 Frequency

🛠 Spatial ETL Tool

🗂 USDOT\_NAD.pyt

🗗 Check Geometry

🗗 Domain Analysis

🗗 Feature Count

🗁 ST (State two letter code)

🗁 ST\_YYMMDD (date of submission)

🗁 PreETL\_QC

🛢 Submitted\_geodatabase.gdb

∴ Source\_Address\_Pts

🖹 ST\_FeatCount.csv

🗁 Result\_QC

🛢 ST\_ETL\_results.gdb

🗁 Flagged

∴ Geometry

∴ Duplicates

∴ Etc…

∴ ST\_postETL

🖺 ST\_YYMMDD\_Report.docx

🛠 ST.fmw