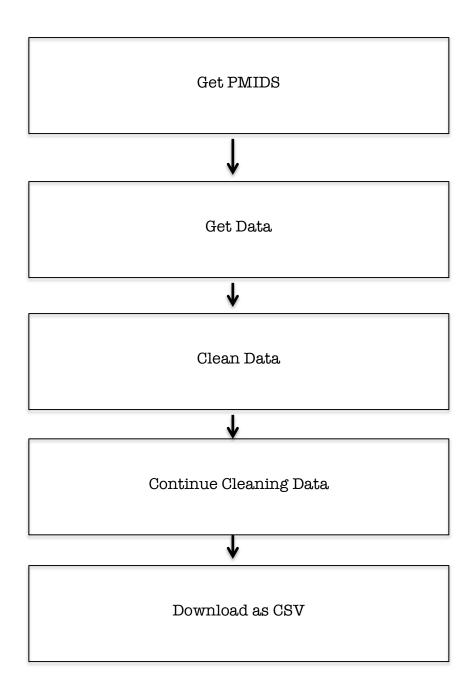
## Section 1: Program Design

This project is divided into three modules: A data scraper module, a SQL module and a data visualization model. The following pages contain the workflow diagrams detailing the structure and function of these modules.

## **Data Scraper**

These are the five main functions performed in the scraper module, each of which has it's own detailed workflow diagram in the subsequent pages:



### **Get PMIDs** User enters search term, max & min dates Loop 1 Biopython Package Entrez searches through Loop 2 PubMed entries that meet search criteria Does the article meet Does the the criteria? article meet Keep Keep the criteria? Searching Searching No Yes No Yes Add 1 to Counter Add PMID to list Are there more Are there articles to more search? articles to search? No Yes Yes No You have Convert your retmax from list to You have value. series to your list of Return for proceed to PMIDs. Loop 2 get data

### **Get Data**

For each PMID in series created in previous step:

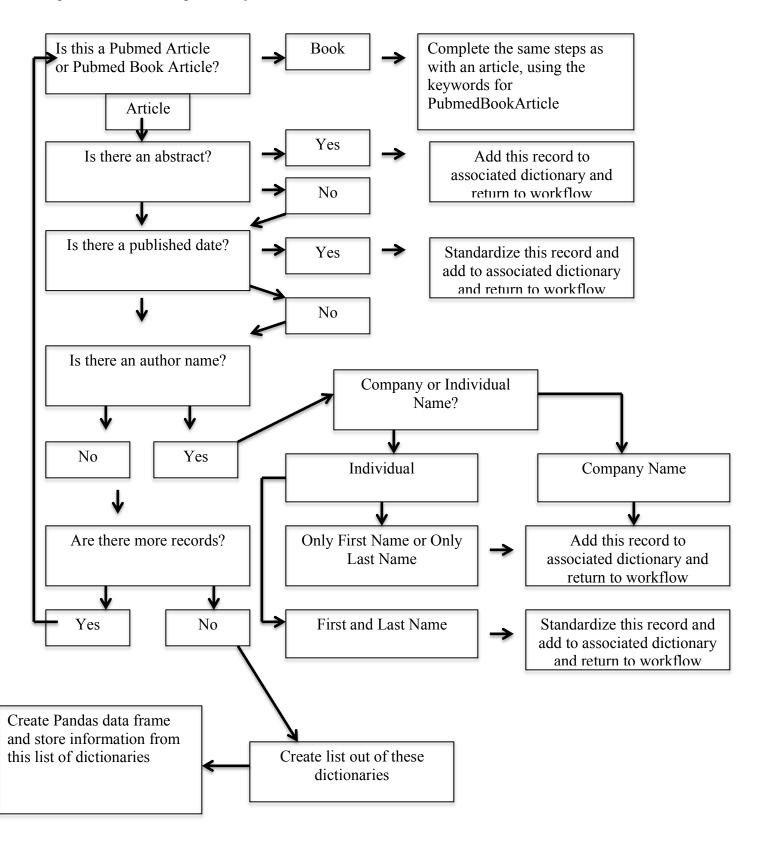
Use Entrez efetch function to return record as handle in xml format

Use Entrez read function to parse XML results into Python dictionary

Append record into dictionary

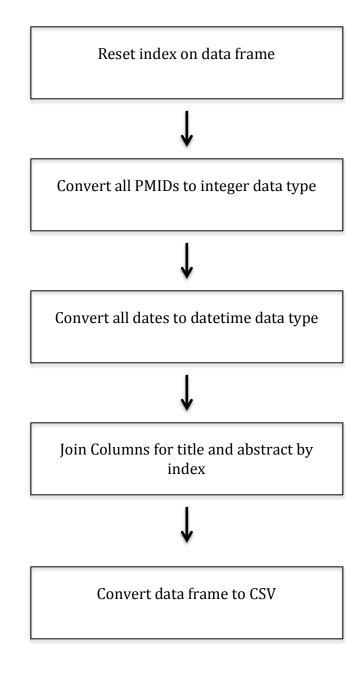
### Clean Data

This function uses a list of dictionaries to store all citation data for the dataset. The following is performed for each previously retrieved record:

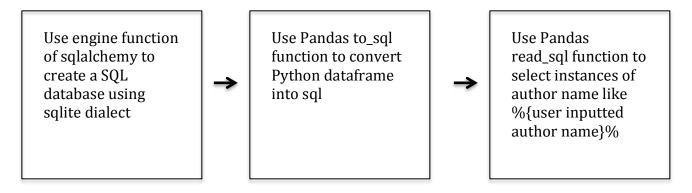


# Continue Cleaning Data & Convert to CSV

Begin with dataframe from previous workflow:



### **SQL** Module



### Visualization Module

Draw Graph

Line

# Create categorical variable for month Count instances of Published date in each month category Allow user to manually select from following graphs to display results

Bar

Both

# **Summary Statistics**

