**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

Get Data

Clean Data

Continue Cleaning Data

Download as CSV

**Get PMIDs**

Biopython Package Entrez searches through PubMed entries that meet search criteria

User enters search term, max & min dates

Yes

Does the article meet the criteria?

Keep

Searching

Yes

Does the article meet the criteria?

Keep

Searching

Add 1 to Counter

Are there

more articles to search?

No

No

No

Yes

You have your retmax value. Return for Loop 2

You have your list of PMIDs.

Convert

from list to series to proceed to get data

Add PMID to list

Are there

more articles to search?

No

Yes

**Loop 2**

**Loop 1**

**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:

Is this a Pubmed Article

or Pubmed Book Article?

Complete the same steps as with an article, using the keywords for PubmedBookArticle

Is there an abstract?

Is there a published date?

Add this record to associated dictionary and return to workflow

Standardize this record and add to associated dictionary

and return to workflow

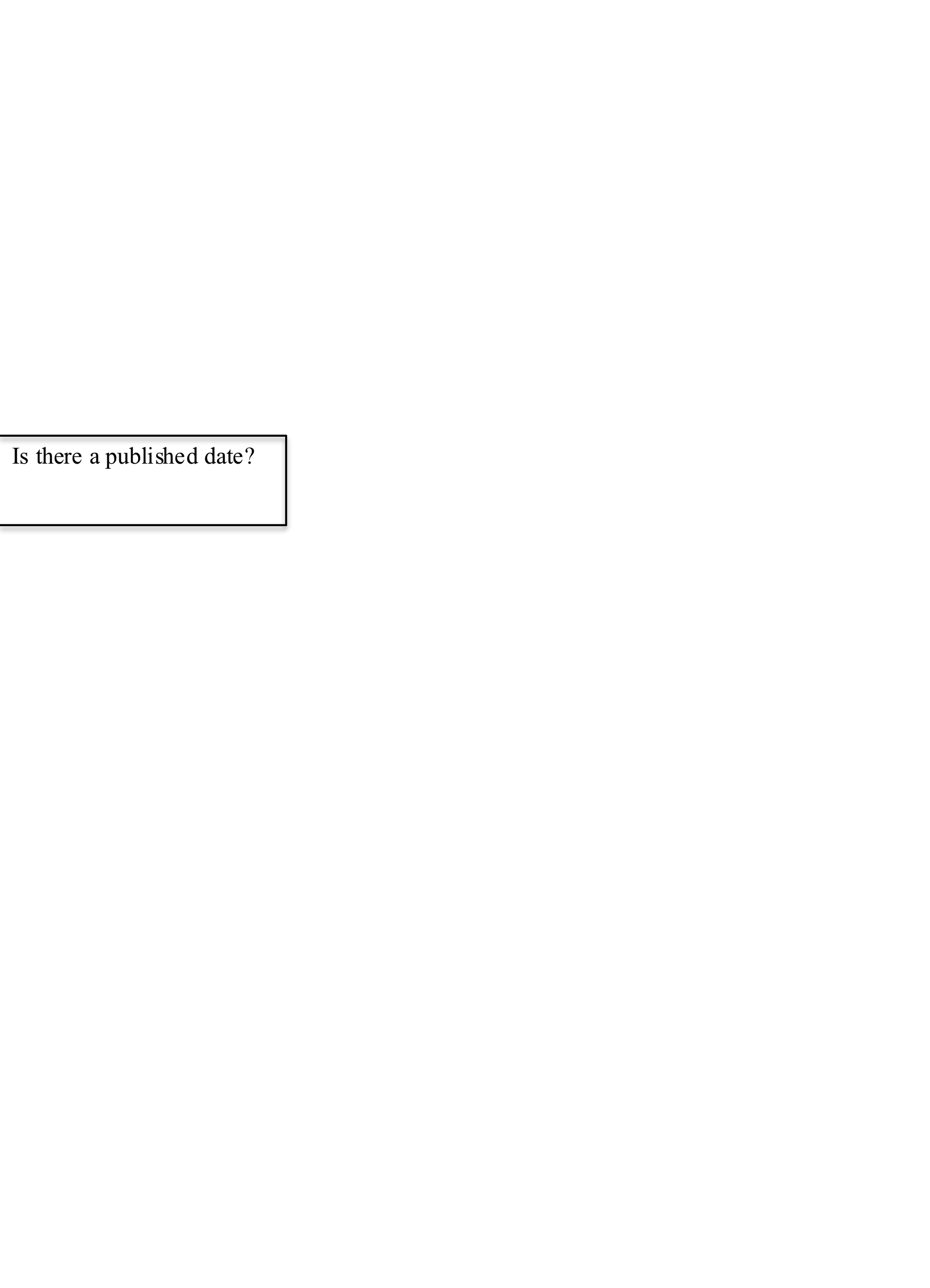
Yes

Yes

No

No

Is there an author name?

Create list out of these dictionaries

Create Pandas data frame and store information from this list of dictionaries

No

Yes

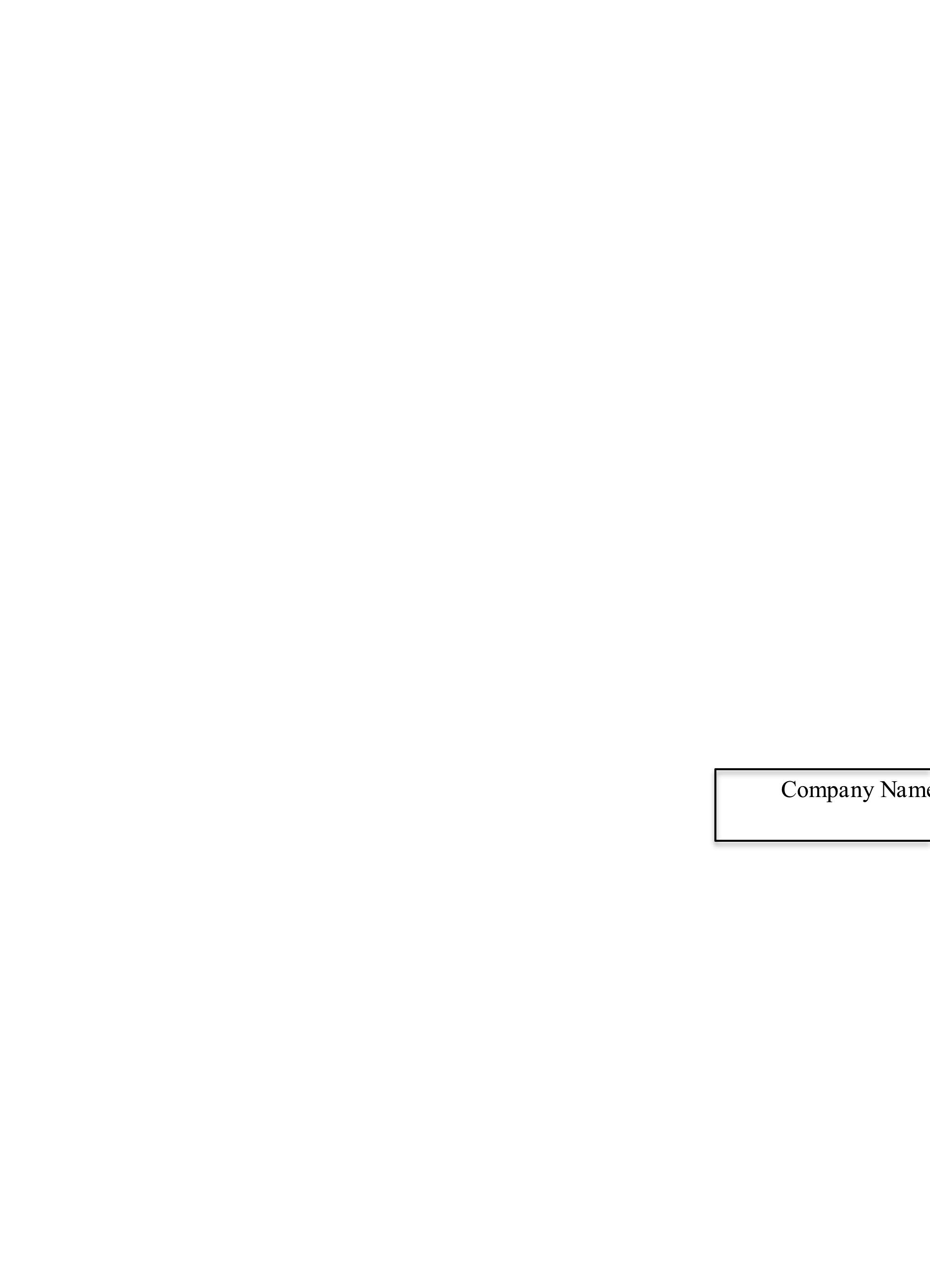
Are there more records?

Yes

No

Company or Individual Name? 

Company Name

Individual Name

Only First Name or Only Last Name

First and Last Name

Add this record to associated dictionary and return to workflow

Book

Article

Standardize this record and add to associated dictionary

and return to workflow

**Continue Cleaning Data & Convert to CSV**

Begin with dataframe from previous workflow:

Convert all dates to datetime data type

Convert all PMIDs to integer data type

Reset index on data frame

Join Columns for title and abstract by index

Convert data frame to CSV

**SQL Module**

Use engine function of sqlalchemy to create a SQL database using sqlite dialect

Use Pandas to\_sql function to convert Python dataframe into sql

Use Pandas read\_sql function to select instances of author name like %{user inputted author name}%

**Visualization Module**

**Draw Graph**

Count instances of Published date in each month category

Create categorical variable for month

Create copy of data frame

Allow user to manually select from following graphs to display results

Line

Bar

Both

**Summary Statistics**

Create copy of data frame

Create categorical variable for month

Count instances of Published date in each month category

Use Pandas to create summary statistics for this newly created distribution