# Lyme Disease Project Update 11 August 2021

## Weekly Progress

Downloaded details of 12000 CrossRef articles from 2010 – 2021

Downloaded details of 9988 Google Scholar articles from 2010 – 2018, 2021 (see issues)

Downloaded details of 1348 PubMed articles from 2010 – 2021

Search terms “lyme disease tick-borne”

Created DataFrames of each source.

Began data cleaning process

Plotted top sources – see plots later in the document.

### Notes about Data Quality

**Abstract Availability**

Google ≈94% (but shortened to around 250 characters)

PubMed ≈97% (perhaps full abstracts, but smaller number of results)

CrossRef ≈13% (some full).

**DOI Availability (Primary candidate for common key on which to merge results)**

Google ≈23%

PubMed ≈92%

CrossRef ≈100%

**URL Availability (for scraping additional abstracts.**

Google ≈96%

PubMed ≈0%

CrossRef ≈100%

## Issues & Mitigation

|  |  |  |
| --- | --- | --- |
| Issue | Mitigation | Status |
| Only 1000 articles returned at a time | Broke down searches year by year  Wrote script to call PoP from command line via Jupyter notebook to automate search and save process (success with CrossRef and PubMed, limited success with Google – see next issue) | CrossRef and PubMed complete.  Google Scholar search completed for 10 of 12 years from 2010 to 2021. |
| Google Scholar has locked me out due to too many searches. Two years (2019 and 2020) missing in google data. | Wait? |  |
| PoP has generated publication names containing an ellipsis character “…” at the beginning or end of shortened titles in Google results.  This inconsistency between titles may affect which publications are listed in the top X of google sources | Option 1: Data cleaning using difflib.getclosematches() and SequenceMatcher to match with full names where possible and update dataframe.  Option 2: Match with corresponding result in CrossRef (by DOI) and update | In progress |
| Results contain some Non-English articles | Apply some filtering criteria | To Do |
| More filtering of results required  Limit to journal articles?  Are all articles from *Ticks and Tick-Borne Diseases* relevant or has name of journal caused all its articles (incl. announcements) to be returned? | Get input from Teo and Stephen | To Do |

Most Common Journals

Graphical user interface

Description automatically generated

### Cross Ref sources: Top 20

Text, letter

Description automatically generated

Chart

Description automatically generated

Chart, funnel chart

Description automatically generated

Data Cleaning

CrossRef

Handful of articles containing the word “announcement” – deleted

Filtered – either article or publication must contain “tick-borne|Tick-Borne|Tick-borne|Lyme|lyme",

Managed to complete google set of

Article has to

Have either a working doi, article url or full url.

Has to have an abstract

Must have lyme or tick-borne in title or

Has to be a journal article

Has to be in English

Has to have a title

Thorough