Towards More Effective Querying of Medical Literature in Alexandria3K

How useful can Alexandria3K be for performing literature reviews?

Bas Verlooy

Supervisors: Diomidis Spinellis, Georgios Gousios

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
January 28, 2024

Name of the student: Bas Verlooy
Final project course: CSE3000 Research Project
Thesis committee: Diomidis Spinellis, Georgios Gousios, Koen Langendoen

Abstract

The Alexandria3K library, a versatile Python-based tool, has been expanded to include the integration of the PubMed dataset, enriching its capabilities in the analysis of scientific papers. Originally supporting major datasets like Crossref and US patents, and smaller yet significant datasets like ORCID. The addition of PubMed enables in-depth analysis of medical papers, with medical specific data and articles not yet in the Crossref dataset. This research focused on validating the integration of PubMed into Alexandria3K. To achieve this, two literature surveys were replicated using the complete PubMed dataset. The first survey involved querying different pathogens in the dataset for three regions. The results were comparable, although some articles were missed by Alexandria3K but two articles were also missed by the original survey. The second survey revolved around software tools used in medical papers. Although fewer articles were found with Alexandria3K the ratio for most tools was still comparable. Although a thorough manual review of all articles could have further refined the reevaluation, time constraints prevented this step. These replicated surveys demonstrate Alexandria3K's potential in conducting literature surveys, underscoring the need for manual validation to complement its capabilities.

1 Introduction

The Alexandria3K library, developed in Python, offers a versatile platform for analyzing a broad choice of datasets related to research papers and patents [14]. It currently supports data extraction from Crossref, Open Researcher and Contributor ID (ORCID), the US Patent (USPTO) dataset, along with several other smaller yet significant datasets.

Alexandria3K has proven effective in tracking the evolution of different types of papers, such as synthesis studies, using simple SQL queries on the databases it creates. This capability is exemplified in Figure 1, which illustrates the fluctuating trends in synthesis research obtained from the Crossref dataset [14]. The methodologies for these analyses are accessible in the examples directory on GitHub, demonstrating the library's utility and transparency.

Building upon these foundations, this study explored the integration of the MEDLINE database into Alexandria3K. Managed by the United States National Library of Medicine, MEDLINE is a comprehensive resource for life sciences and biomedical information [12]. PubMed, a free search engine, broadens the scope of MEDLINE by including additional data such as pre-1966 citations and articles beyond life sciences journals. As the National Library of Medicine states, "MEDLINE is the largest subset of PubMed" [10]. The inclusion of PubMed data in Alexandria3K, including articles lacking a *doi* and thus not in the CrossRef dataset,[1] enriches the library's dataset with detailed medical information, like Medical Subject Headings (MeSH), examined chemicals, and abstracts in languages other than English.

To demonstrate the enhanced utility of Alexandria3K with the integration of PubMed, two medical literature surveys were reanalyzed using the library. These reevaluations, compared to the original survey results, could reveal Alexandria3K's potential in simplifying article discovery and analysis in PubMed. The extensive use of PubMed data for bibliometric analysis, as indicated by Ioannidis (2016) with over 325,000 items tagged as systematic reviews and meta-analyses, underscores the importance of an efficient analysis tool like Alexandria3K [5].

The methodology for conducting literature surveys varies, as an analysis of 15 bibliometric analyses on PubMed showed. These ranged from exclusive use of PubMed's search engine to employing additional tools like the R language [1; 16], tools using the PubMed database [7] or visual analysis software [6; 13]. Unlike existing web-based tools for PubMed that primarily present results as article lists [8], Alexandria3K offers options for Python, SQL, and a command line interface, facilitating deeper analysis with raw data. This unique feature sets Alexandria3K apart from other tools that either require chaining multiple tools for in-depth analysis or focus on a limited subset of papers. A similar tool is pubmed.mineR [11], this tool allows the user to search through abstracts and automatically extract keywords. With further development, this could also be developed for Alexandria3K as the data is already available and could be a great extension for PubMed as well as the other datasets.

One survey reevaluated using Alexandria3K focused on the occurrence of pathogens listed by the WHO in research, particularly in the Gulf Region and Bahrain [2]. This reanalysis using Alexandria3K's capabilities of searching through PubMed paper abstracts, keywords, and MeSH terms highlights the library's efficiency, though it is constrained to the content of available abstracts.

Additionally, the study tested Alexandria3K's effectiveness against the combined use of Google Sheets and SPSS as employed by Masuadi et al. (2021) in their analysis of the PubMed database [9]. This comparison aimed to determine if Alexandria3K could serve as a standalone tool for similar analyses, providing results comparable to traditional methods.

This research was guided by the primary question: "How can Alexandria3K help perform literature surveys on medi-

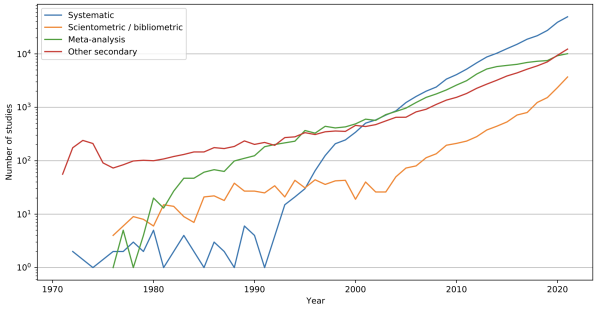


Figure 1: Synthesis studies over time [14].

[1] https://www.doi.org/

cal papers?". To address this the first subquestion was how data from PubMed could be incorporated in Alexandria3K. A good starting point for this were the datasets already integrated in Alexandria3K.

Sub-questions related to the analysis of literature surveys help determine Alexandria3K's effectiveness. These include assessing the quantity of results compared to original surveys. Lastly, the quality of the articles found was analyzed for the first reevaluation by manually comparing all articles found for Bahrain.

2 Methodology

This section is divided into two distinct parts: the first focuses on the integration of PubMed into Alexandria3K, and the second assesses the practical utility of Alexandria3K through two distinct case studies.

2.1 Relational view of PubMed

Alexandria3K already supports eight diverse data sources, ranging in size from 25 MB to over 1 TB for the expanded Crossref dataset. The PubMed dataset, encompassing the entirety of the MEDLINE database [15], is on a similar scale to the Crossref dataset. Smaller datasets within Alexandria3K are directly downloadable via their URLs, whereas larger datasets like PubMed require pre-download due to their substantial size.

To integrate PubMed, parallel to the approaches for Crossref, ORCID, and USPTO datasets, the initial step was sourcing the PubMed files. These files, hosted on an FTP server, range in size from 2.6 MB to 86 MB compressed, totaling 45 GB compressed and 702 GB uncompressed.[2]

The XML format of the PubMed data facilitated the reuse of methods applied to the USPTO dataset, such as extracting data from attributes and looping through XML arrays. Unlike the USPTO dataset, however, all PubMed files are stored in a single directory, simplifying importing the data as the file name has no added benefit to the dataset. The Document Type Definition (DTD) file from PubMed was utilized to identify all attributes within the files, aiding in the creation of a robust data schema. Obsolete fields, as indicated in the DTD file and accompanying documentation, were omitted to maintain consistency with other datasets in Alexandria3K.

Given the overlap of some tables with those in the Crossref dataset and the presence of unique articles in PubMed, the decision was made to include all overlapping tables fully, which also enables the use of the PubMed dataset without standalone.

The data can be extracted into a SQLite database similar to the other datasets as documented in the documentation of Alexandria3K. Listing 1 shows how it can be imported through the command line, similarly, it can also be done through Python.

[2]https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/README.txt

```
python3 bin/a3k populate database.db pubmed
    "./pubmed_data/"
```

Listing 1: Populating PubMed database from the command line

2.2 Reevaluation of literature survey

Post-implementation of PubMed in Alexandria3K, its usefulness was explored through the replication of existing literature surveys. The focus was on quantitative reviews, which focus on the amount of articles found for certain keywords or metadata of articles.

One such survey [2] involved searching for WHO-listed pathogens in PubMed, with a specific focus on the Gulf Region, especially Bahrain. A list of twelve pathogens was queried and compared across three regions. The entire PubMed dataset was used to ensure comprehensive and comparable results. The following tables were used:

- *Articles*: Listing unique article identifiers and metadata.
- *Abstracts*: Full abstracts for keyword and country searches.
- *Keywords*: Including publisher-provided keywords.
- *MeSH*: Covering Medical Subject Headings relevant to pathogens and countries.

Initial analysis using these tables individually missed several articles, the subsequent analysis combined the article, abstract, and MeSH tables for keyword searches. The keywords table was disregarded as it showed no added benefit and did slow down queries due to more joins needed between the tables. The need to extract the country data from the MeSH and abstract table arose because a simple SQL query revealed a mere three articles originating from Bahrain-based journals as can be seen in Listing 2. In the end, the approach was to find the pathogens in the MeSH table, there should be a mention of *resistance* in the MeSH, abstract, or title. Lastly, if there also was a criterion for a specific region, one of those countries for that region should also be present in one of those three tables.

```
SELECT COUNT(*) FROM pubmed_articles
   WHERE journal_country = 'Bahrain';
-> 3
```

Listing 2: SQL query to fetch articles published in Bahrain

To enhance query performance, indexes were created in the SQLite tables, facilitating faster query execution through B-trees without storing data [4]. The initial search for pathogens was stored in a new table to prevent the same query from having to run for all different regions.

Furthermore, the full-text search (FTS) feature of SQLite was leveraged for efficient word searches in abstracts and titles, saving time compared to regular SQL text searches. FTS automatically indexes the table and allows for multiple keyword matches in a single statement, a comparison in queries

can be seen in Listings 3 and 4. It should be noted that for both options the query is case insensitive.

```
SELECT DISTINCT(article_id) FROM pubmed_abstracts
    WHERE text LIKE '%Bahrain%' OR text LIKE
        '%Saudi Arabia%';
```

Listing 3: SQL query to get articles which mention Bahrain or Saudi Arabia in their abstract

```
SELECT DISTINCT(article_id) FROM pubmed_abstracts
    WHERE text MATCH '"Bahrain" OR "Saudi Arabia"';
```

Listing 4: SQL query to get articles which mention Bahrain or Saudi Arabia in their abstract using FTS

2.3 Alexandria3K as statistical software

Previous research by Masuadi et al. (2021) [9] involved querying PubMed for statistical tool usage in publications, using Google Sheets and SPSS to analyze the results. The articles were filtered by year and statistical software used, allowing for the analysis of tool usage over time.

Replicating this research with Alexandria3K involved similar steps, where the article and the abstract table were used for searching the software tools. Again the keywords table was excluded due to its negligible improvement and increased performance cost due to joining tables. All abstracts from articles from 1997, 2007, and 2017 were indexed in a new FTS table for quick access. While the original study included a manual review of articles for context validation, due to time constraints this step was omitted in the Alexandria3K analysis.

After having all articles and abstracts from the relevant years a count was made for each year for articles where one or more software tools were mentioned in the title or abstract. The first results showed a lot less results than expected, after this for each software tool synonyms and full names were also included in the search query. Subsequently, for each software tool, the database was queried for the amount of times mentioned. Since an article could use multiple software tools the total percentage of tools used is higher than 100%.

3 Contributions

This section details the contributions made during this research, structured into two subsections. The first part highlights the coding enhancements to the Alexandria3K library, and the second part explores the scientific applications of the contributions.

3.1 Coding Contributions to Alexandria3K

Throughout this research, several enhancements were made to the Alexandria3K library. The most notable of these is the incorporation of the PubMed dataset, which introduced 21 new tables covering a wide range of data from publication details to the chemicals used in research papers. This integration facilitates both vertical and horizontal slicing of the dataset. Vertical slicing allows users to select specific tables or columns, while horizontal slicing offers filtering or sampling options before data ingestion. Both options allow for the option to populate the database faster and vertical slicing gives the option to only retrieve the data needed for the research.

Efforts were made to minimize code redundancy by repurposing existing code from other datasets and moving that code to generic files for shared functionalities. New helper functions were also devised to address PubMed-specific requirements. Lastly, all new code is tested according to the standards set by the existing code.

Additional minor yet impactful enhancements include the introduction of a progress bar feature. Enabled by the *-d progress_bar* option, this feature visually indicates the progress of dataset loading, an improvement over the previous method that only displayed the file currently being loaded. Another advancement is the ability to define column data types, particularly beneficial for XML-formatted data like PubMed, where previously all columns defaulted to string type. This was not the case with JSON files as the data type can be read while reading the file for that format. The addition of data types allows for filtering by year ranges, for example, all articles published after 2000, while previously it was only possible to filter by a specific year with string comparison for XML datasets.

3.2 Validating PubMed

To validate PubMed's utility, two literature surveys were reanalyzed using Alexandria3K. The first step involved populating a database with the complete PubMed dataset, which totals 156 GB and approximately 1.74 billion rows. The focus was on a quantitative literature review that counted the number of published articles for specific keywords. It required multiple iterations of improvements to achieve results comparable to the original.

The first survey reevaluation involved searching the PubMed database for twelve different pathogens identified by the WHO as having antibiotic resistance. The original survey's methodology was replicated, with searches conducted in the abstracts, keywords, and MeSH tables. The review was divided into three stages: a global search, a targeted search within the Gulf Cooperation Council (GCC), and a focused search regarding Bahrain. For the articles found or not found regarding Bahrain, a comparison was made between the differences. For each article, it was validated why it was not included or if it should have been included in the original survey.

Lastly, a statistical analysis was recreated by querying the PubMed database trying to follow the steps listed in the original paper. It was not possible to validate all articles found for the matched keywords due to time constraints.

4 Results

This section presents the results in two parts: the analysis of the PubMed dataset integration into Alexandria3K, and the reevaluation of the literature surveys.

4.1 Dataset

Table 1 showcases a detailed breakdown of the total rows per table in the PubMed dataset as of January 2024. Revealing insights such as the fact that 25% of the articles lack a *doi*, making them unlinkable to the Crossref dataset. Each article in the dataset may have multiple abstracts, each labeled to indicate its purpose, and some articles include abstracts and titles in languages other than English. There's even one article that contains abstracts in twelve languages [3] regarding the agricultural measures in the EU with 34 authors from different countries.

The full schema of PubMed, including 21 new tables, is detailed in Appendix A. The *pubmed_articles* table can be linked with the *works* table from the Crossref dataset through the *doi*. The author table can also be linked with the ORCID table through the ORCID identifier.

A comparison of PubMed and Crossref datasets, based on the number of rows per table, is shown in Table 2, highlighting the sizeable differences between the two datasets.The data used is from previous research by Spionellis (2023) [14].

It can be seen that the total amount of records from Crossref is only 45% bigger than PubMed although the amount of publications is 277% bigger. This is mainly because PubMed contains 21 tables, while Crossref contains 10 tables.

4.2 Survey reevaluation

The complete PubMed dataset was utilized for the experiments. Due to the low number of articles from the GCC and Bahrain, using a subset of the dataset would have skewed the results.

Since no machine learning is involved, each experiment is only conducted once, although multiple iterations of queries were necessary to get the desired result. The queries used are also publicly available in the GitHub repository.

As noted in the methodology there were four tables considered for searching the PubMed database: articles, abstracts, keywords, and MeSH. The keywords table was left out as it had no added benefit and a combination of the other tables was used in the end. The experiment's outcome can be seen in table 3, while the list of pathogens searched for can be found in Appendix B.1.

For all but two pathogens the amount of articles found was bigger using Alexandria3K when not restricted to any region. For one pathogen the difference was even 8.5 times bigger than the original result. The results for only the GCC region were already a lot more aligned. For the three pathogens which showed a lot more articles found with Alexandria3K 50% was manually analyzed to confirm that the results were plausible. After this analysis, no articles were found that should not have been identified by Alexandria3K.

For the Bahrain region, there was a mismatch of five articles not found by Alexandria3K and one article not found in the original survey. For all articles that were found in the original survey, but not by Alexandria3K, and the other way around, it was analyzed why it was not included. The full list of these articles is included in Appendix B.2.

For the five articles not found by Alexandria3K two times Bahrain was not mentioned. Two times the pathogen MeSH listed was not the same as mentioned in the list from the WHO. Finally, the last article was not found as it did not have a PubMed ID or *doi* and was thus not in the Alexandria3K database and was impossible to find.

Alexandria3K identified two articles that the original literature survey had missed. One of those articles had been included for a different pathogen, and according to the methodology of the original survey, an article could be included for multiple pathogens. Accordingly, this article was included in two pathogens in the reevaluation. The other one could have been included in the original analysis as it fulfilled all criteria as manual assessment confirmed.

With regards to the FTS table, the speed comparison was significant, as can be seen in the comparison in Listing 5. Both queries had the same data and produced the same results. The FTS table did take up a few extra GB of storage space but did perform 3350 times as fast compared to the query without FTS.

```
SELECT COUNT(*) from pubmed_abstracts WHERE text
    LIKE "%shigella%";
    -> 67 seconds
SELECT COUNT(*) from fts_abstracts WHERE text
    MATCH "shigella";
    -> 0.02 seconds
```

Listing 5: Comparison of two queries of which one utilises FTS, averaged over 10 runs.

4.3 Statistical reproduction

With the search for articles with one or more of the software tools listed 7841 articles were found. In comparison in the original analysis, 10 596 articles were found. Of those, 9427 were included as the remaining articles lacked an abstract or accessible full text. A direct year-by-year comparison was hindered as the original report did not specify the annual breakdown of these figures. A plausible reason for the discrepancy in the total count might be attributed to the original survey's methodology, which included an examination of the methods section within each article, data not available in Alexandria3K's PubMed dataset.

In Figure 2 the differences can be seen between the original analysis and the analysis performed by Alexandria3K, both overall and across three distinct years. A notable observation is the overrepresentation of articles mentioning SAS and *Review Manager* in the Alexandria3K dataset. This overrepresentation is likely due to the lack of manual verification of these terms within the articles, compounded by the fact that SAS has multiple interpretations in the biomedical field, at least five of which are documented.[3]

On the other hand, the *R Project* presented a different challenge, the original survey reported using several synonyms but did not specify them. Hence, synonyms such as *R software*, *R package*, *R language*, and CRAN were utilized. Interestingly, the SPSS package was consistently underrepresented in the Alexandria3K findings across all years. This underrepresentation remained unexplained, even after extending the

[3] https://en.wikipedia.org/wiki/SAS

Table 1: Amount of rows per table for PubMed

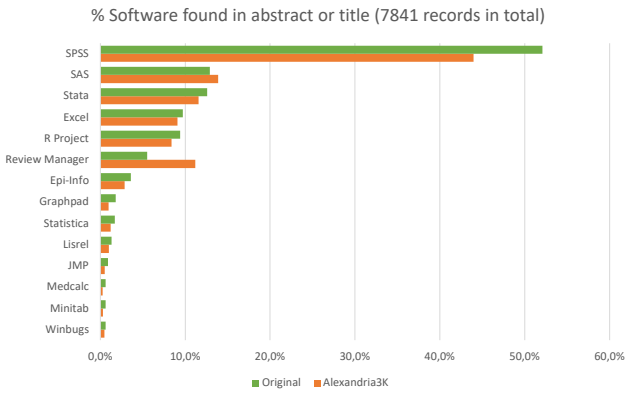
Table	# Records
Articles (publications)	36 555 430
Articles with DOI	27 670 023
Articles with abstracts	20 145 494
Articles with abstracts available in original language	204 934
Articles with title available in original language	3 755 829
Articles without title in English	78 641
Authors	160 853 266
Author affiliations	96 301 037
Abstracts	45 302 379
Abstract groups in another language	251 815
Abstracts in another language	406 002
Unique abstract labels	30 221
Chemicals	136 512 825
Comment corrections	2 881 759
Data banks	475 915
Data bank accessions	435 988
Grants	12 223 915
History	136 512 825
Investigators	5 081 919
Investigator affiliations	153 171
Keywords	50 008 791
Unique keywords	6 664 244
MeSH	321 519 467
MeSH supplements	209 153
Publication types	50 008 791
References	352 881 053
Reference articles	325 573 532

Table 2: Comparison in size between PubMed and CrossRef [14]. For the comparison, the amount of rows from Crossref is divided by the amount of rows from PubMed. Rows in thousands.

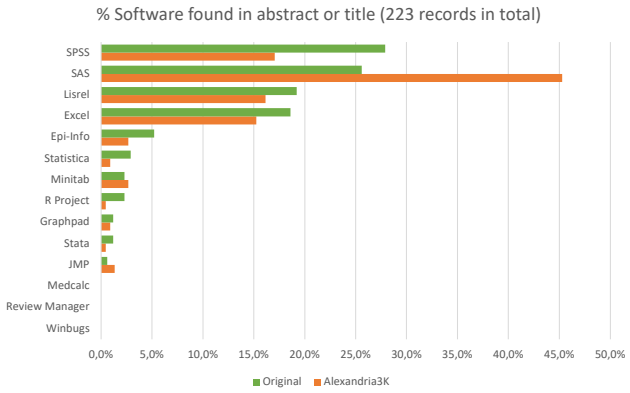
Table	PubMed/10^3	Crossref/10^3	Compared
Total records	1 740 843	2 531 227	1.45x
Publications	36 555	134 048	3.77x
Authors	160 853	359 556	2.24x
References	352 881	1 748 421	4,95x
References with doi	10 279	1 255 033	122.09x
Distinct doi references	5885	59 127	10.04x

Table 3: Comparison between the original survey vs. Alexandria3K database used to search for the same pathogens.

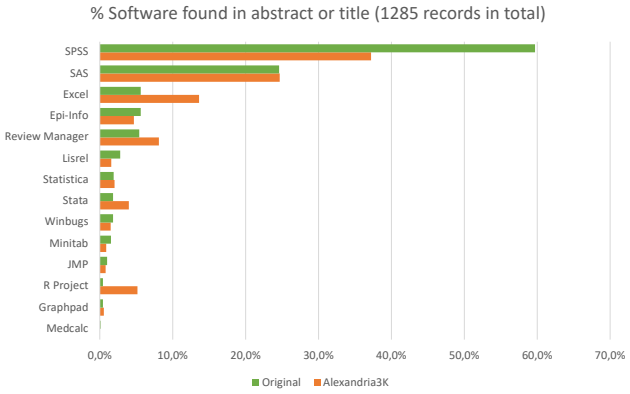
Region	Source	1	2	3	4	5	6	7	8	9	10	11	12
Globally	Original	768	467	1169	2622	28 023	271	404	886	219	164	1693	234
	Alexandria3K	2424	7740	4822	1833	21 955	2755	1484	7527	1760	5343	2191	1148
GCC	Original	30	9	16	12	278	3	4	19	1	5	21	11
	Alexandria3K	31	16	26	6	80	10	8	26	2	36	14	10
Bahrain	Original	2	-	4	1	3	1	1	-	1	-	1	-
	Alexandria3K	2	-	1	-	2	1	1	-	1	1	1	-



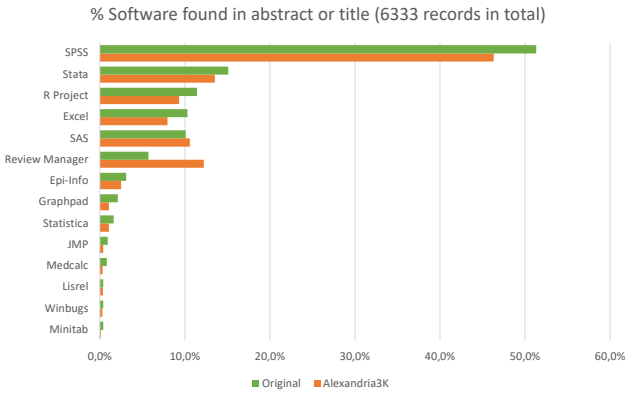
(a) Overall differences



(b) Differences in 1997



(c) Differences in 2007



(d) Differences in 2017

Figure 2: Total and yearly differences between original analysis vs. Alexandria3K

search to include the full name "Statistical Product and Service Solutions".

The last step of the statistical analysis was not performed as it meant going over each article to find the study design. As the number of articles amounted to a couple thousand, this was deemed not feasible in the time given.

5 Responsible Research

In addressing the aspects of responsible research for this research project, it is essential to highlight the measures taken to ensure both ethical integrity and reproducibility.

Concerning reproducibility, all code developed for this project is publicly available on GitHub.[4] This allows other researchers to replicate the methods and results shown in this paper, enabling a collaborative and open research environment. The contributions shown in this paper can be looked back on by searching for the pull requests done by the author.

The inclusion of an */examples/* directory on GitHub plays a crucial role in ensuring the reproducibility of this research. This directory contains the code and queries that are used in the graphs and results shown and discussed in this paper. It serves a dual purpose: firstly, it provides a real-world context for the theoretical aspects of the library, and secondly, it acts as a step-by-step guide for replicating the results presented. By exploring these examples, readers can gain a deeper understanding of how the library functions in various scenarios.

On the ethical front, this research strictly utilizes publicly available data that is free to download. The use of such data ensures adherence to ethical standards, avoiding any potential issues related to data privacy or unauthorized use of proprietary information. By opting for openly accessible sources, the research aligns with the principles of ethical data usage, reinforcing the integrity of the study.

6 Discussion

The PubMed dataset is one of the larger datasets for Alexandria3K, although still more than 1.5 times as small in amount of records and 3.7 times as small in amount of publications. This size difference is partially attributed to the specific focus of PubMed on medical sciences, necessitating numerous medical-specific tables that are not present in more generalized datasets like Crossref. Crossref, catering to a broader range of scientific fields, does not delve as deeply into field-specific data structures as PubMed does. In the end, the data schema was formed by the DTD file, documentation, and the already existing datasets.

While implementing the new PubMed dataset a starter guide might have been useful for implementing a new datasets. It could have been useful to have a list of steps to begin looking in the source code and an overview of how to start developing a new dataset. At the moment of writing instructions on an installation environment are given and the required CI/CD tools are listed. A guide on how to start developing a new dataset could include an overview of how a dataset is created using the aspw[5] library and which Python classes are required for each component of a dataset.

[4] https://github.com/dspinellis/alexandria3k

[5] https://pypi.org/project/apsw/

The experience of implementing PubMed was generally positive, despite challenges encountered with nested tables and XML file formatting. These issues were resolved through effective debugging. A key learning point for future dataset integrations is the importance of sourcing data from the correct source that allows comprehensive file downloads. During the project, the PubMed site was blocked due to not adhering to the terms set by the site. This required the use of a new account that wasn't blocked to continue research and development, while using the FTP server would have prevented this issue from the start.

Reevaluating the literature survey using the local Alexandria3K database revealed both advantages and disadvantages. While web-based tools offer ease of access without the need for installation or data downloading, they cannot match the convenience of having a complete dataset locally available, despite the considerable storage requirements (157 GB for the complete PubMed dataset). Populating the database can be time-intensive, but this process can be managed by running it overnight.

A downside of only having the abstracts available instead of the full texts was noticed when performing the statistical analysis. In the original search query, the statistical tools were searched for in the title, abstract, and full text. This caused fewer articles to be found than the original analysis. This might have been the reason for the difference of 1586 missing from the second reevaluation.

One of the benefits of Alexandria3K is that results can be stored in between steps. For both validation of the intermediate results and speed improvement, this can be useful. Furthermore, with Alexandria3K all data available to an article can be used after filtering. After searching the MeSH for certain keywords it was still possible to filter the abstracts by other keywords as the tables could still be joined. This allowed for the queries to be improved for the first analysis without having to rerun all initial queries.

Another benefit is that the data can be queried using SQL and Python and thus allows for complex queries to be built. This allows for the aggregation of data to for example create graphs grouped by year.

However, the statistical analysis underscored the need for content review within each article, a task outside the current scope of Alexandria3K. For comprehensive analysis, data would need to be exported to software like Excel for article labeling or filtering. Nonetheless, the simplicity and flexibility of SQL queries, further enhanced by FTS for rapid text field queries, underscore the robustness and adaptability of Alexandria3K in bibliometric analysis.

7 Conclusions and Future Work

In this paper, it is shown that PubMed can be integrated into the Alexandria3K library with the help of the Document Type Definition for defining a data schema. The already implemented datasets guided in how to integrate PubMed into the existing framework. The choice was made to include overlapping fields with other datasets, among other things this means that PubMed could be used standalone.

To show the capabilities of Alexandria3K a literature survey was reevaluated as well as a statistical analysis. The literature survey regarding pathogens was reevaluated by searching the Medical Subject Headings for a list of pathogens. These were then further filtered through a geographical scope, firstly globally, then for the Gulf Cooperation Council, and lastly only for Bahrain. Out of the 14 articles from Bahrain found in the original survey, five were not found with Alexandria3K. There were different three different reasons for this, firstly twice Bahrain was not mentioned, two times the pathogen from the WHO was different than the one listed in the article and finally, one article was not indexed in PubMed and impossible to find. Two articles were found with Alexandria3K and not included in the original. After manual confirmation, the conclusion was that these two articles could have been included in the original.

The statistical analysis replicated using Alexandria3K centered on identifying common software tools used in medical research. Alexandria3K detected fewer articles compared to the original study (7841 vs. 9427), partly because the original analysis included searches in the methodology sections of articles, a data aspect not available in Alexandria3K's dataset. Despite this, the results for most software tools were closely aligned with the original study, barring a few exceptions due to an ambiguous abbreviation and unexplained underrepresentation for one tool.

Overall, the reevaluations demonstrated Alexandria3K's effectiveness in producing results comparable to traditional literature surveys. While manual review of articles, a step undertaken in the original surveys, could potentially enhance accuracy, time constraints prevented this in the study.

Looking ahead, further research could involve reevaluating more literature surveys using Alexandria3K to solidify its utility in aiding researchers. Such studies could extend not only to additional PubMed datasets but also to those from Crossref, broadening the scope and applicability of Alexandria3K in diverse research domains. This future work would contribute significantly to validating and enhancing Alexandria3K as a versatile tool for academic and scientific research.

A PubMed schema

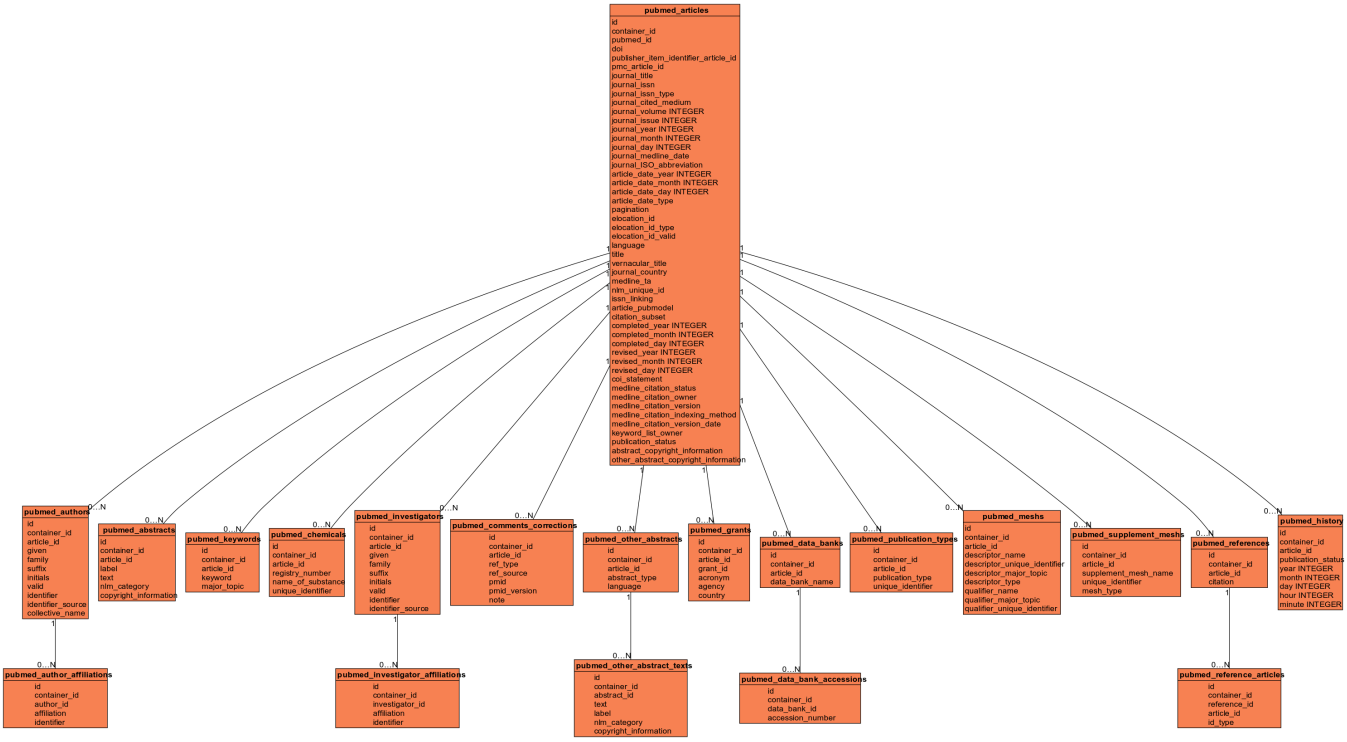


Figure 3: PubMed schema

B Pathogens

B.1 Pathogen list WHO

Table 4: Pathogens published by WHO and results from evaluated Survey [2]

Number	Pathogen	type	Global	GCC	Bahrain
1	Acinetobacter baumannii	Critical	768	30	2
2	Pseudomonas aeruginosa	Critical	467	9	-
3	Enterobacteriaceae	Critical	1169	16	-
4	Enterococcus faecium	High	2622	12	1
5	Staphylococcus aureus	High	28 023	278	3
6	Helicobacter pylori	High	271	3	1
7	Campylobacter	High	404	4	1
8	Salmonella	High	886	19	-
9	Neisseria gonorrhea	High	219	1	1
10	Streptococcus pneumonia	Medium	164	5	-
11	Haemophilus influenzae	Medium	1693	21	1
12	Shigella	Medium	234	11	-

B.2 Pathogen differences

Table 5: Pathogen differences between Alexandria3K and original survey [2]

PubMed ID	Pathogen	Not found by	Reason
22118027	3	Alexandria3K	No correct pathogen MeSH
25312123	3	Alexandria3K	No correct pathogen MeSH
No PubMed/doi	3	Alexandria3K	Not indexed in PubMed, thus impossible to find
24231491	5	Alexandria3K	No Bahrain mentioned
17314416	7	Alexandria3K	No Bahrain mentioned
18219143	7	Original	Could be included
27048582	10	Original	Could be included

References

[1] ASIRI, F. Y., KRUGER, E., AND TENNANT, M. Global dental publications in pubmed databases between 2009 and 2019-a bibliometric analysis. *Molecules (Basel, Switzerland) 25* (10 2020).

[2] ASOKAN, G. V., RAMADHAN, T., AHMED, E., AND SANAD, H. WHO global priority pathogens list: A bibliometric analysis of Medline-PubMed for knowledge mobilization to infection prevention and control practices in bahrain. *Oman Med J 34*, 3 (May 2019), 184–193.

[3] COLE, L. J., KLEIJN, D., DICKS, L. V., STOUT, J. C., POTTS, S. G., ALBRECHT, M., BALZAN, M. V., BARTOMEUS, I., BEBELI, P. J., BEVK, D., BIESMEIJER, J. C., CHLEBO, R., DAUTARTĖ, A., EMMANOUIL, N., HARTFIELD, C., HOLLAND, J. M., HOLZSCHUH, A., KNOBEN, N. T. J., KOVÁCS-HOSTYÁNSZKI, A., MANDELIK, Y., PANOU, H., PAXTON, R. J., PETANIDOU, T., DE CARVALHO, M. A. A. P., RUNDLÖF, M., SARTHOU, J.-P., STAVRINIDES, M. C., SUSO, M. J., SZENTGYÖRGYI, H., VAISSIÈRE, B. E., VARNAVA, A., VILÀ, M., ZEMECKIS, R., AND SCHEPER, J. A critical analysis of the potential for EU common agricultural policy measures to support wild pollinators on farmland. *The Journal of Applied Ecology 57* (4 2020), 681–694.

[4] GAFFNEY, K. P., PRAMMER, M., BRASFIELD, L., HIPP, D. R., KENNEDY, D., AND PATEL, J. M. Sqlite: Past, present, and future. *Proc. VLDB Endow. 15*, 12 (aug 2022), 3535–3547.

[5] IOANNIDIS, J. P. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly 94* (9 2016), 485–514.

[6] LI, S., WANG, H., ZHENG, H., LI, N., SUN, C., MENG, X., ZHENG, W., WANG, K., QIN, H., GAO, W., AND SHEN, Z. Bibliometric analysis of pediatric liver transplantation research in pubmed from 2014 to 2018. *Medical science monitor : international medical journal of experimental and clinical research 26* (6 2020), e922517.

[7] LIAO, K.-Y., WANG, Y.-H., LI, H.-C., CHEN, T.-J., AND HWANG, S.-J. Covid-19 publications in family medicine journals in 2020: A pubmed-based bibliometric analysis. *International journal of environmental research and public health 18* (7 2021).

[8] LU, Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database 2011* (1 2011), baq036–baq036.

[9] MASUADI, E., MOHAMUD, M., ALMUTAIRI, M., ALSUNAIDI, A., ALSWAYED, A. K., AND ALDHAFEERI, O. F. Trends in the usage of statistical software and their associated study designs in health sciences research: A bibliometric analysis. *Cureus 13* (1 2021), e12639.

[10] NATIONAL LIBRARY OF MEDICINE. MEDLINE, PubMed, and PMC (PubMed central): How are they different?

[11] RANI, J., SHAH, A. R., AND RAMACHANDRAN, S. PubMed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. *Journal of Biosciences 40* (10 2015), 671–682.

[12] RETHLEFSEN, M. L. MEDLINE: A guide to effective searching in PubMed and other interfaces. *Journal of the Medical Library Association 95* (4 2007), 212–213.

[13] SHIAO, C.-C., WU, J.-T., CHU, Y.-C., TANG, Y.-H., HUANG, L., AND LAI, H.-Y. Bibliometric analysis of the top 100 most-cited articles on video laryngoscope from 2011 to 2022. *Journal of the Chinese Medical Association : JCMA 86* (10 2023), 902–910.

[14] SPINELLIS, D. Open reproducible scientometric research with Alexandria3k. *PLoS ONE 18*, 11 (Nov. 2023), e0294946.

[15] UNIVERSITY OF MEDICINE, J. H. Expert searching.

[16] XU, S., FU, Y., XU, D., HAN, S., WU, M., JU, X., LIU, M., HUANG, D.-S., AND GUAN, P. Mapping research trends of medications for multidrug-resistant pulmonary tuberculosis based on the co-occurrence of specific semantic types in the mesh tree: A bibliometric and visualization-based analysis of pubmed literature (1966-2020). *Drug design, development and therapy 17* (2023), 2035–2049.