

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Graph-based Tool for Exploring PubMed Knowledge Base

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Bottoni, S., Trombetta, A., Bertini, F., Montesi, D., Bonin, F., Pascale, A., et al. (2023). Graph-based Tool for Exploring PubMed Knowledge Base. New York : IEEE Computer Society [10.1109/ICDE55515.2023.00280].

Availability:

This version is available at: <https://hdl.handle.net/11585/939860> since: 2023-08-30

Published:

DOI: <http://doi.org/10.1109/ICDE55515.2023.00280>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

S. Bottoni *et al.*, "Graph-based Tool for Exploring PubMed Knowledge Base," *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, Anaheim, CA, USA, 2023, pp. 3611-3614.

The final published version is available online at:
<https://dx.doi.org/10.1109/ICDE55515.2023.00280>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

Graph-based Tool for Exploring PubMed Knowledge Base

Simone Bottoni¹

University of Insubria

Varese, Italy

sbottoni@uninsubria.it

Alberto Trombetta¹

University of Insubria

Varese, Italy

alberto.trombetta@uninsubria.it

Flavio Bertini²

University of Parma

Parma, Italy

flavio.bertini@unipr.it

Danilo Montesi³

University of Bologna

Bologna, Italy

danilo.montesi@smartdata.cs.unibo.it

Francesca Bonin⁴

IBM Research Europe

Dublin, Ireland

fbonin@ie.ibm.com

Alessandra Pascale⁴

IBM Research Europe

Dublin, Ireland

apascale@ie.ibm.com

Martin Gleize⁴

IBM Research Europe

Dublin, Ireland

martin.gleize@ie.ibm.com

Pierpaolo Tommasi⁴

IBM Research Europe

Dublin, Ireland

ptommasi@ie.ibm.com

Abstract—Studies have shown that data retrieval and visualization tools can help health professionals to improve their understanding and communication with patients, their relationship with stakeholders, and their decision-making process. However, not many efforts have been made in this direction. In this paper, we present a prototype system for the indexing, annotation, and visualization of the PubMed knowledge base to enable the search and retrieval of health-related evidence. The proposed tool builds and keeps updated an enriched graph based on PubMed articles associating them with concepts extracted from the Unified Medical Language System (UMLS) Metathesaurus. Moreover, it allows a full-text search and graph-based navigation and supports an overview of concepts and related publications. The proposed architecture enables scale-up thanks to its containerized nature and parallelization capabilities. The code is open-source under the Apache V2 license.

Index Terms—Indexing, annotation, graph-based knowledge visualisation, PubMed knowledge base

I. INTRODUCTION

Recent events such as the COVID-19 pandemic have shown us the importance of leveraging health data and informatics systems for scientific discovery, evidence generation and management, and delivery of care services. Many such services rely on evidence-based decision-making, policies, and budget allocation. A robust evidence base constitutes one of the pillars of the scientific approach to providing healthcare services. To build such a base, access to and management evidence is essential. Many efforts in the medical community have been devoted to using Artificial Intelligence (AI) technologies in healthcare to extract evidence automatically. However, we believe evidence retrieval and visualization have been given a lower priority. A keyword search on PubMed (a free search engine for references and abstracts on life sciences and biomedical topics) for “AI health” in the last five years returns 61k results, while a search for “health data visualization” only returns 2.9k results. On the other hand, we believe that most stakeholders in the field would benefit from a tool that would help them retrieve and visualize evidence presented in scientific literature [1]. This would help, among others, the

topical debate on explainability, which is one of the main aspects causing patient apprehension regarding the use of AI in healthcare [2]. Some work has been focusing on temporal visualization of electronic health records [3], while others on Excel-based data visualization for healthcare students [4]. In this context, we have also seen the development of PubMed derivatives such as GoPubMed [5] or SemanticMedline [6]. GoPubMed is able to extract Gene Ontology (GO) [7] terms from the PubMed publications abstract, and groups the them according to the GO terms. Instead Semantic MEDLINE provides to the users predictions based on UMLS concepts extracted from PubMed.

We present a graph-based tool for indexing, annotating, and visualizing PubMed knowledge base, an open source system providing a graph-based explorative rendering of the PubMed information, which is augmented with medical concept annotations. The graph-based tool enables non-IT experts to access information quickly, while medical concept annotations allow the creation of a linked knowledge base of evidence literature. Specifically, the developed tool consists of the following capabilities:

- processes and indexes the annual baseline and daily updates of PubMed¹;
- annotates articles with medical concepts using the UMLS compendium [8];
- keeps updated the repository used by the tool on a daily basis;
- exposes an interface allowing health personnel searching and graph-based data rendering.

Our tool aims to facilitate access to published evidence and improve the discovery of evidence of health factors within scientific works. It was built to handle large workloads through a containerized structure which allows the system to easily scale and increase the data processing capacity. The system

¹We refer here only to the publicly available information provided by PubMed, i.e., abstract and available information about authors, affiliation, etc.

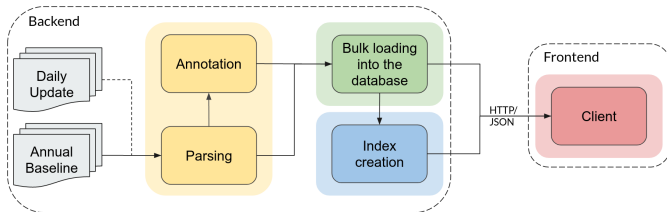


Fig. 1: Workflow of the system.

has been made open source under Apache V2 license².

Our tool was developed within the context of Social Determinants of Health (SDoH) studies. SDoHs have been the hearth of many initiatives, including nutritional programs in food safety and housing interventions tackling homelessness [9]. One of the previous studies' [10] [11] bottlenecks is the acquisition of scientific articles in a scalable, constant, and up-to-date manner, as well as in being able to perform natural language queries for SDoH related terms. Our system can help researchers retrieve information and understand the relationship between SDoH and health conditions.

In the next sections we describe the dataset and the set of Unified Medical Language System (UMLS) taxonomy used for annotation (Section II and Section III), the overall workflow in Section IV, and the detailed description of the architecture in Section V. Finally, conclusions and future works are reported in Section VI.

II. PUBMED DATA

PubMed is one of the most significant scientific repositories of biomedical and life sciences literature. It contains more than 33 million in citations and abstracts from life science journals, newspapers, magazines, and newsletters dating back to 1809.

PubMed is managed by the U.S. National Library of Medicine (NLM). Every year the NLM publishes the annual baseline that contains the whole PubMed citation records (i.e., articles) in XML format up to the current year. The NLM also publishes updated files daily, including new, revised, or deleted articles. Each XML file contains articles with a specific structure that must conform to the PubMed Document Type Definition (DTD). The structure includes a collection of basic information about an article, such as an id, called PubMed ID (PMID), the title, the abstract, the journal, the publication date, the authors with the affiliations list, and the citations list. It can also contain more specific biological information like genetic symbols and chemical substances.

III. UNIFIED MEDICAL LANGUAGE SYSTEM

UMLS is a set of vocabulary, classification systems, and coding standards (implemented in files and software) used to enhance the effectiveness and interoperability of biomedical information systems [8]. It includes different knowledge sources, including the Metathesaurus, a set of hierarchies, definitions, and other relationships and attributes.

UMLS Metathesaurus [12] is an extensive, multi-lingual vocabulary database containing information about biomedical and health-related concepts, their various names, and their relationships. It is organized by concept or meaning, connecting different names over a unique and permanent Concept Unique Identifier (CUI). Different tools use UMLS Metathesaurus to extract specific terminologies or ontologies from text. One of the most widely used programs is MetaMap [13] and its lite version MetaMapLite, a customizable and faster version of MetaMap [14]. Using Natural Language Processing techniques, these named entity recognition tools are used to map biomedical text to the UMLS Metathesaurus concepts.

IV. WORKFLOW

The overall objective of the tool is to provide an easier and faster way to explore the evidence in clinical literature. To reach this goal, we first need to index the PubMed repository and ensure that the index stays updated daily, and second, expose all the available information in a user-friendly way, enhancing them with the UMLS annotations. In this section, we present the workflow and its main phases (Figure 1); we explain how PubMed data are indexed, processed, and annotated.

The proposed tool first downloads the annual baseline and daily update XML files from the PubMed repository. It currently downloads, processes, and then deletes a user-defined number of files at a time to ease system pressure, especially on the processors and disk. Once a PubMed file is downloaded, it starts the processing workflow comprising three phases: parsing, annotation, and collection/indexing.

a) Parsing: In this phase, an XML parser takes input from an XML PubMed file. It extracts all the related metadata, converting the information presented in Section II to feed the relational database used by our tool.

b) Annotation: The annotation process consists in analyzing the input text and finding a match with the UMLS Metathesaurus concepts or meanings, presented in Section III, describing specific articles' characteristics.

c) Collection and indexing: The parsed and annotated (articles and related concepts) PubMed baseline data is collected from the respective phases and then stored in a relational database. To improve performance and lower access time, a further index - optimized for textual data - is built upon a subset of interest (e.g., we currently focus on data regarding authors, articles, and concepts rather than on the history of articles).

As seen in Section II, NLM releases files containing updated information on the articles every day; therefore, we have a recurrent job looking for updates on the articles and processing them through the presented workflow. We then store all the updated information in the database incrementally, keeping track of history.

Although it is not part of the workflow, the annotations and all relevant metadata are exposed externally through REST Application Programming Interface (API) and queried through a Web User Interface (UI) (as we will see in Section V).

²The system is open source and accessible for deployment at the following link: <https://github.com/SimoneBottoni/PubMedKnowledgeGraph>

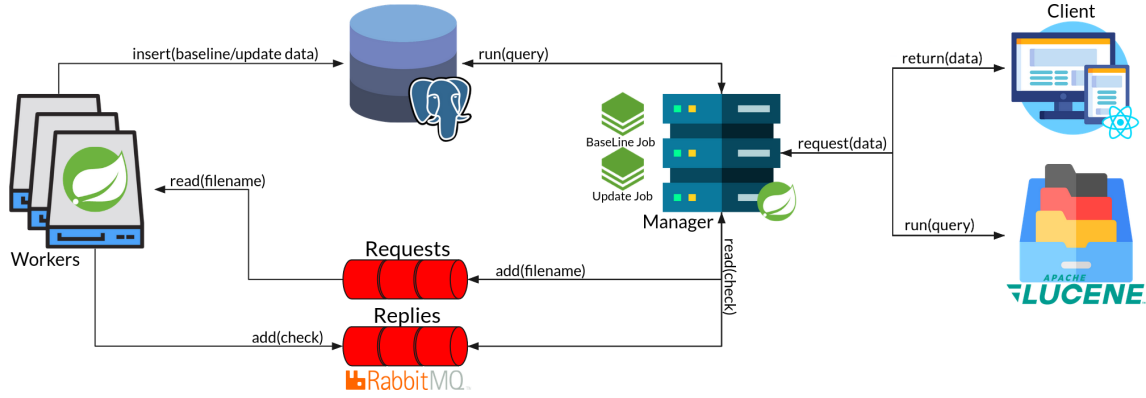


Fig. 2: Architecture of the system.

V. ARCHITECTURE

In this section, we present a high-level overview of the architecture (Figure 2). The architecture has three main components: a *manager* that orchestrates the interactions, a set of *workers* that concurrently processes the PubMed files, and a *web UI* to query and visualize the data.

A. Manager

The *manager* is the core component of the system. It is written using the Spring Boot framework [15] and governs all the main functionalities of the system: initialising, orchestrating all the jobs involved in analysing PubMed data, maintaining indexes, and fulfilling clients' queries.

a) Initialization: During the bootstrap, the *manager* ensures the integrity of the underlying database schema and instantiates two messaging queues (i.e., requests and replies queues) for the communications with the *workers*. For our purposes, we used PostgreSQL as a database engine and RabbitMQ as a queues manager.

b) Job Orchestration: Once the bootstrap is completed, the *manager* can execute the workflow over PubMed data as seen in Section IV. To process the PubMed files, the *manager* scatters the files to be processed using remote partitioning jobs [16], an abstraction that allows to easily distribute tasks across the different *workers* over the network. We defined two types of remote partitioning jobs: *BaseLine job* to process the PubMed annual baseline through the BaseLine messages, and *Update job* to process the daily updates through the Update messages.

- BaseLine messages are issued in the requests queue by the *manager*, using the PubMed files identified after bootstrapping. Details on how the *workers* process these messages are available in the next Section V-B. The *manager* keeps track of the *BaseLine job* progress and, once completed, it enables the *Update job*.
- Update messages are issued at a later stage. Every day, at a given user-defined time, the *manager* awakes to handle the *Update job* for the daily updates released by PubMed. When awoken, the *manager* checks the

PubMed website for the newly updated files and then issues Update messages in the requests queue.

c) Index Maintenance: Once the BaseLine and Update jobs are completed, the *manager* updates index of the most relevant data. This index, built using Hibernate Search [17] and relying on Apache Lucene [18], covers a pre-defined list of attributes (e.g., title, abstract, authors, tags, etc).

d) Fulfilling Clients Queries: Last duty of the *manager* is to handle the requests coming from the web clients. A full-text search (exploiting the indexes) and an overview of the relationships among articles and concepts are exposed through a set of RESTful API using JSON as the data-interchange format.

B. Worker(s)

Firstly, a *worker* parses the information regarding the articles using Java Sax parser. Then, the *worker* annotates each article exploiting MetaMapLite's API to generate a list of keywords. MetaMapLite's API takes a text document as input, matches it with the UMLS Metathesaurus, and outputs a list of relevant concepts that have been identified. Each concept is retrieved in the MetaMap Indexing (MMI) format, which includes, among others, the keyword (also referred to as CUI), the corresponding name in natural language (also referred to as *preferred term*), a score, and the trigger information. The score represents the relevance of the UMLS concept according to the MMI ranking function [19]; the trigger information includes some information regarding the word of the input text that triggered the keyword and its location in the input text. The *worker* passes to the MetaMapLite's API the parsed title and abstract of each article. Lastly, in the collection phase, the *worker* saves the articles' data and the keywords in the database. Once all the phases are covered, the *worker* notifies the *manager* through a status message on the replies queue.

C. Web UI

We paired our system with a modern multi-page React web application to query and explore the PubMed knowledge graph, exploiting the REST APIs provided by the *manager*. We split our *web UI* into two components: one to enable a full-text search over the indexed data, and one to explore

the relationships among concepts and articles in a 3D graph fashion.

a) *Full-Text Search*: The full-text search component allows users to query data using natural language and obtain a list of articles that match the query. Each entry of the result represents a single article, showing essential information such as the title, the author(s), the journal, the abstract, and the matched UMLS concepts (i.e., a list of pairs CUI and preferred term). Also, each entry of the result has a detailed view that shows all the available information regarding the selected article. It includes primary article information, the list of the article's references with the PubMed Central (PMC), Digital Object Identifier (DOI), and PMID links, and all the information related to the article's status in the PubMed repository (i.e., the owner, the last revision date, the copyright information, and other available details). The full-text search includes a temporal filter to screen the results, for instance, it is possible to set a time interval to show the results belonging to the desired period.

b) *3D Graph Exploration*: The 3D graph exploration component provides a graphical overview of the concepts, the relationships among them, and the related publications. It helps understand the connections between different articles and the UMLS concepts, with further explanations about discovered relationships or information about concepts. The 3D graph exploration component offers different options to search for data. For instance, it is possible to search articles by the PMID (the article identifier), the CUI (the UMLS concept identifier), or via a UMLS preferred term. The resulting entities are visualized as a 3D graph in the form of nodes and edges. To render the graph, we used *react-force-graph*, a React library [20] displaying 3-dimensional space graph data structure rendered using a force-directed iterative layout. In particular, each node can represent an article, an author, a journal, or a UMLS concept. The center node of the graph is always the searched entity. Every node and edge allows visualising of further details about the entity or the relationship between entities.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a graph-based tool for the indexing, annotation, and visualisation of PubMed knowledge base that facilitates the retrieval of evidence and enriched concepts. In particular, the tool associates specific UMLS Metathesaurus concepts to the PubMed scientific literature to simplify the search for evidence and possible links to health issues. We provide a set of accessible APIs to query the processed data and a modern multi-page web application to search and visualize the PubMed knowledge base. A screencast of the demo of the proposed tool is available at the following link (<https://youtu.be/wkILp-IKBUy>). The tool is able to scale up providing the possibility to instantiate new workers based on the workload that must be processed. However, further evaluations must be conducted to test and understand the actual performance of the tool to process the whole PubMed repository (i.e., the necessary time) to provide technologies

requirements in terms of computational capabilities. Future studies will focus on the improvement of the data exploration and visualization tools and a formal usability study of them. These studies should be performed with the help of the users who are going to use the system (e.g., doctors and researchers) that can provide reviews and suggestions on the utility and usability of the presented tools.

REFERENCES

- [1] S. Park, B. Bekemeier, and A. Flaxman, "Understanding data use and preference of data visualization for public health professionals: A qualitative study," *Public Health Nursing*, vol. 38, 02 2021.
- [2] J. P. Richardson, C. Smith, S. Curtis, S. Watson, X. Zhu, B. Barry, and R. R. Sharp, "Patient apprehensions about the use of artificial intelligence in healthcare," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–6, 2021.
- [3] N. Martign  ne, T. Balcaen, G. Bouzill  , M. Calafiore, J.-B. Beuscart, A. Lamer, B. Legrand, G. Fich  ur, and E. Chazard, "Heimdall, a computer program for electronic health records data visualization," *Studies in health technology and informatics*, vol. 270, pp. 247–251, 2020.
- [4] F. LaPolla, "Excel for data visualization in academic health sciences libraries: a qualitative case study," *Journal of the Medical Library Association*, vol. 108, 01 2020.
- [5] A. Doms and M. Schroeder, "Gopubmed: Exploring pubmed with the gene ontology," *Nucleic acids research*, vol. 33, pp. W783–6, 08 2005.
- [6] H. Kilicoglu, M. Fiszman, A. Rodriguez, D. Shin, A. Ripple, and T. Rindfleisch, "Semantic medline: A web application for managing the results of pubmed searches," *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, 11 2007.
- [7] M. Harris, J. Deegan, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. Rubin, J. Blake, C. Bult, M. Dolan, H. Drabkin, J. Eppig, D. Hill, L. Ni, and R. White, "The gene ontology (go) database and informatics resource," *Nucleic Acids Res*, vol. 32, pp. 258–261, 02 2004.
- [8] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.
- [9] *Beyond health care: the role of social determinants in promoting health and health equity*, 2018.
- [10] Y. Park, N. Mulligan, M. G. amd Morten Kristiansen, and J. H. Bettencourt-Silva, "Discovering associations between social determinants and health outcomes: Merging knowledge graphs from literature and electronic health data," in *AMIA 2021, American Medical Informatics Association Annual Symposium, Virtual Event, USA*. AMIA, 2021.
- [11] M. Gleize, N. Mulligan, A. Di Bari, and J. H. Bettencourt-Silva, "Social determinant trends of covid-19: An analysis using knowledge graphs from published evidence and online trends," in *MIE*, 2021, pp. 744–748.
- [12] P. L. Schuyler, W. T. Hole, M. S. Tuttle, and D. D. Sherertz, "The umls metathesaurus: representing different views of biomedical concepts," *Bulletin of the Medical Library Association*, vol. 81, no. 2, p. 217, 1993.
- [13] A. R. Aronson, "Metamap: Mapping text to the umls metathesaurus," *Bethesda, MD: NLM, NIH, DHHS*, vol. 1, p. 26, 2006.
- [14] D. Demner-Fushman, W. J. Rogers, and A. R. Aronson, "MetaMap Lite: an evaluation of a new Java implementation of MetaMap," *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 841–844, 01 2017. [Online]. Available: <https://doi.org/10.1093/jamia/ocw177>
- [15] Spring-Projects, "spring-boot," <https://github.com/spring-projects/spring-boot>, 2018.
- [16] SpringProjects, "spring-batch," <https://github.com/spring-projects/spring-batch>, 2008.
- [17] Hibernate, "hibernate-search," <https://github.com/hibernate/hibernate-search>, 2010.
- [18] A. Bialecki, R. Muir, and G. Ingersoll, "Apache lucene 4," in *OSIR@SIGIR*, 2012.
- [19] A. R. Aronson, "The mmi ranking function," Available in the website: <https://ii.nlm.nih.gov/MTI/Details/mmi.shtml>, 1997.
- [20] V. Asturiano, "react-force-graph," <https://github.com/vasturiano/react-force-graph>, 2018.