Building the COVID-19 Portal By Integrating Literature, Clinical Trials, and Knowledge Graphs

Aoshen Wan
School of Information
University of Texas at Austin
Austin, TX 78701
Email: orsonwan@utexas.edu

Yan Zhan
School of Information
University of Texas at Austin
Austin, TX 78701
Email: zhanyan@utexas.edu

Sanjana Tripathi
School of Information
University of Texas at Austin
Austin, TX 78701
Email: sanjana@utexas.edu

Jiarong Yang
School of Information
University of Texas at Austin
Austin, TX 78701
Email: jiarong@utexas.edu

Mona Sachdev
School of Information
University of Texas at Austin
Austin, TX 78701
Email: mona.sachdev@utexas.edu

Shreya Paithankar
Department of Pediatrics and Human Development
Michigan State University
Grand Rapids, MI 49503
Email: paithank@msu.edu

Joel Duerksen
Data2Discovery
Greater Orlando, Florida
Email: joel@d2discovery.com

Bin Chen
Department of Pediatrics and Human Development
Michigan State University
Grand Rapids, Michigan 49503
Email: chenbi12@msu.edu

Ying Ding
School of Information
University of Texas at Austin
Austin, Texas 78701
Email: ying.ding@ischool.utexas.edu

***Abstract*—The outbreak of COVID-19 has a severe impact on our families, communities, and businesses. Researchers, practitioners, and administrators need a tool to help them digest this enormous amount of knowledge to address various scientific questions related to COVID-19. With CORD-19 dataset, this paper showcases the COVID-19 portal to portray the research profiles of scientists, bio entities (e.g., gene, drug, disease), and institutions based on the integration of CORD-19 research literature, COVID-19 related clinical trials, PubMed knowledge graph, and the drug discovery knowledge graph. This portal provides the following profiles related to COVID-19: 1) the profile of a research scientist with his/her COVID-19 related publications and clinical trials with tweets amount; 2) the profile of a bio entity which could be a gene, a drug, or a disease with articles and clinical trials; and 3) the profile of an institution with papers authored by researchers from this institution.**

I. Introduction

The outbreak of COVID-19 has a severe impact on our families, communities, and businesses. Scientists are working around the clock to find cure and save lives. Efficient communication are essential for us to have a clear picture on the current status of scientific endeavors to facilitate COVID-19 research and collaboration.

This paper showcases the COVID-19 portal to portray the research profiles of scientists, bio entities (e.g., gene, drug, disease), and institutions based on the integration of CORD-19 research literature, COVID-19 related clinical trials, PubMed knowledge graph[1], and the drug discovery knowledge graph[2]. This portal provides the following profiles related to COVID-19: 1) the profile of a research scientist with his/her COVID-19 related publications and clinical trials with tweets amount; 2) the profile of a bio entity which could be a gene, a drug, or a disease with articles and clinical trials; and 3) the profile of an institution with papers authored by researchers from this institution.

II. Methodology

A. Data

This portal is based on the integration of data from literature, clinical trials and knowledge graphs. Literature includes the CORD-19 dataset which contains the published articles or preprints related to COVID-19. Up till now, 410,682 COVID-19 articles are included in our portal. Clinical trials related to COVID-19 are manually collected from the clinicaltrails.gov website. So far, we have 269 clinical trials in our portal. We normalize and extend our literature and clinical trails using two knowledge graphs: PubMed Knowledge Graph[1] and Drug Discovery Knowledge Graph[2].

B. Architecture of the Portal

COVID-19 portal is implemented by Go and Vue Framework. To setup the database across different data sources, we use PubMed ID to obtain tweet amount from Altmetric and bio entities from Pubtator. To reduce response latency of the portal, we separate authors, bio entities, institutions, and papers into

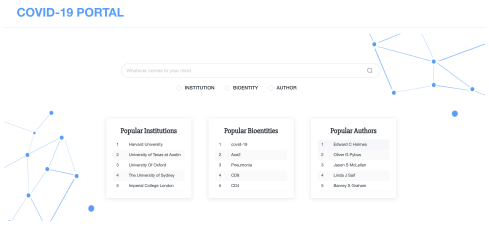

Fig. 1. Front page of the COVID-19 portal

different tables to avoid complex SQL query and limit the response size maximum to 4,000 papers for each query. Figure 1 shows the front page of the portal which highlights the most popular institutions, bio entities, and authors based on the CORD dataset.

III. Result

Our COVID-19 portal is hosted on the web server at School of Information, University of Texas at Austin[1].

A. Front Page

The front page (see Figure 1) was designed in a simple and user-friendly way. The most popular institutions, bio entities and authors are highlighted here that each of them can be clicked which will led to their own profiling page. Auto complete has been enabled when a user is typing an institution, an author, or a bio entity and the relevant auto complete will show up.

B. Profiling COVID-19 Scientists

Author disambiguration is critical. In our prior work of PubMed Knowledge Graph[1], we have used the MapAffil 2016 dataset which contains PubMed authors' affiliations including cities and their geocodes. For those who are the new authors and never published any articles in PubMed before March 2020, we use Semantic Scholar to identify their author's identity and obtain the corresponding affiliation data. For the author profiling page (see Figure 2), the left side is author's contact information and list of his/her publications and clinical trials ranked by publication year or the number of tweets. The right side includes the extracted bio entities from author's publications, co-authors, and the author's yearly publication trend. The author profiling can: 1) raise the scientific community's awareness of who is working on what, with the goal of facilitating potential collaborations; and 2) show the international or inter-regional collaborations of scientific teams and inspire potential collaborations.

C. Profiling COVID-19 Bio Entities

We use PubTator to extract bio entities from the CORD-19 dataset including gene, disease, and drug. Bio entity profiling page have nearly the same layout and sections as the scientists profiling pages. In the bio entity profiling page, we include a bio entity graph including related bio entities to current bio

[1]https://suitclub.ischool.utexas.edu/COVID19/

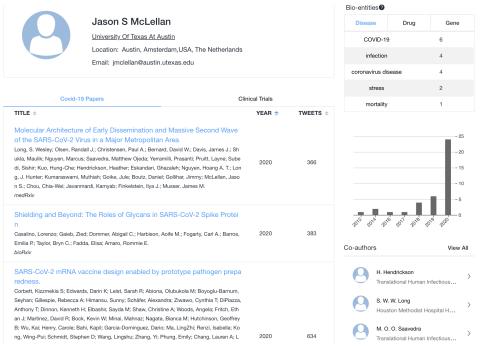

Fig. 2. Author profile with papers, bio-entity and coauthors

entity. For bio entity graph, we connect the CORD-19 dataset with the PubMed Knowledge Graph covering 29 million PubMed articles using BioBert and drug discovery knowledge graph developed by Data2Discovery[2]. This drug discovery knowledge graph integrates data from ChEMBL, PubChem, ExplorEnz, DisGeNET, Disbiome, reactome, UniProt Consortium, neXtProt, TCRD, EMBL, SIDER, stitch, NSIDES, Brown AS and Patel CJ repoDB, NCBI and BgEE. Our portal includes 7 relationships and extracts top 30 related bio entities based on edge score provided by Data2Discovery.

D. Profiling COVID-19 Institutions

For the institution profiling page, besides all the sections in the scientist profiling page, we also include the bio entity word cloud which is related to the papers from current institution. All authors, bio entities, and institutions on each page can be clicked and users will be directed to their corresponding profile pages.

IV. Conclusion

With timely information about the latest scientific development from publications and clinical trials, this COVID-19 portal provides important profiling pages for each author, bio entity, and institution. It establishes the unique summary of the scientific development of COVID-19 based on research profiling of authors, bio entities, and institutions. The same infrastructure can be easily extended to different document corpora, such as the PubMed articles, or AI articles.

Acknowledgment

This project is funded by NSF RAPID (2028717), and Suit Endowment Fund Mary R. Boyvey Dean's Excellence Fund from School of Information at University of Texas at Austin.

References

[1] Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., Rousseau, J.F., Li, X., Xu, W., Torvik, I. V., Bu, Y., Chen, C., Ebeid, I.A., Li, D., Ding, Y. (2020) Building a PubMed Knowledge Graph, Scientific Data, 7, 205.

[2] Gao, Z., Fu, G., Ouyang, C., Tsutsui, S., Liu, X., Yang, J., Gessner, C., Foote, B., Wild, D., Ding, Y., Yu, Q. (2019). Edge2vec: Representation learning using edge semantics for biomedical knowledge discovery. BMC Bioinformatics, 20:306.