COVID-19 Portal: Profiling Researchers, Bio-entities, and Institutions*

Aoshen Wan, Changpeng Tong, Yan Zhan, Sanjana Tripathi, Jiarong Yang, Mona Sachdev, Shreya Paithankar, Joel Duerksen, Bin Chen, and Ying Ding

[1] the University of Texas at Austin, Austin TX 78701, USA
[2] Michigan State University, Grand Rapids MI 49503, USA
[3] Data2Discovery, Greater Orlando, FL, USA

Abstract. The outbreak of COVID-19 has a severe impact on our families, communities, and businesses. White House released the COVID-19 literature dataset (called CORD-19 dataset) which has grown exponentially into a gigantic collection of over 500,000 articles. 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. 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 which can be ranked by year or by the number of tweets; 2) the profile of a bio entity which could be a gene, a drug, or a disease with articles and clinical trials mentioned this bio entity; and 3) the profile of an institution with papers authored by researchers from this institution.

Keywords: COVID-19 · Research profiling · Biological entities · Research activities · Portal design.

1 Introduction

The outbreak of COVID-19 has a severe impact on our families, communities, and businesses. Economies in many countries have plummeted. Scientists are working around the clock to find cure and save lives. Many scientists put the pause button on their own research and switch gears to focus on finding solutions to fight against the COVID-19 pandemic. In early March 2020, White House released the COVID-19 literature dataset [1] with 29,000 COVID-19 related articles. Within one year, this dataset has grown exponentially into a gigantic dataset with over 500,000 COVID-19 related articles in March 2021. Efficient communication of the scientific development and effective measurement

* Funded by NSF RAPID (2028717) and Suit Endowment Fund Mary R. Boyvey Dean's Excellence Fund.

of active scientists, teams, and institutions are essential for us to have a clear picture on the current status of scientific endeavors related to COVID-19. It can facilitate scientists to find collaborators, share resources, and take advantages of collective intelligence to fight against COVID-19. 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 [2], and the drug discovery knowledge graph [3]. 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 which can be ranked by year or by the number of tweets; 2) the profile of a bio entity which could be a gene, a drug, or a disease with articles and clinical trials mentioned this bio entity; and 3) the profile of an institution with papers authored by researchers from this institution.

2 Related Work

Dong, Du, and Gardner [4] developed a real-time COVID-19 dashboard to report the critical statistics about confirmed cases, total deaths, and total recovered. This dashboard provides researchers, public health decision makers, and the public the latest information to track the outbreak. While informative and timely, this dashboard does not cover scientific development activities related to COVID-19. COVID-19 Primer [5] is a COVID-19 portal gathering related articles and news which are classified into a dozen of research topics. It also provides daily summary, emerging topics, and top researchers. But this portal is developed by the company called Primer. It is unclear how the backend data has been collected. For their top ranked researchers, they seem not be able to disambiguate author names. This portal clearly could not provide the research profiles for authors, bio entities, and institutions. COVID Authors [6] is a portal created by the Weber Lab at Harvard Medical School. This portal is more relevant to our portal. It can only do the profiling for authors, but not for bio entities and institutions. They have some keyword profiling, but most of the keywords are not bio entities. Our portal [7] is the central point to provide high quality profiling for researchers, bio entities, and institutions based on the CORD-19 dataset.

3 Methodology

3.1 Bio Entity Extraction

We use PubTator to extract bio entities from the CORD-19 dataset including gene, disease, and drug. PubTator has implemented multiple competition-winning text-mining approaches to automatically identify key biological entities [8]. It can guarantee state-of-the-art performance on biocuration. We have evaluated MetaMap [9], Amazon Comprehend Medical [10], and BioBert [11]. We

have selected PubTator because it is easy to use and provides API to make the update smoothly.

3.2 Author Name Disambiguation

Disambiguating author names is critical. We obtain highquality disambiguated author names from the PubMed literature through the integration of the following two datasets: 1) Author-ity: The Author-ity database uses a variety of information about authors and publications to determine whether two or more instances of the same name (or of highly similar names) on different papers actually represent the same person. According to the author name disambiguation evaluation, the F1 score of Author-ity is 98.16%, which is by far the highest accuracy result that we have observed; 2) Semantic Scholar: Contributors of Semantic Scholar train a supervised binary classifier for merging pairs of author instances and use it to incrementally create author clusters. According to the evaluation, the F1 score of Semantic Scholar is 96.94%. We combined both datasets and collected all the unique authors from PubMed to form PubMed Knowledge Graph [2]. By connecting the CORD-19 dataset with the PubMed Knowledge Graph, we are able to disambiguate author names in the CORD-19 dataset.

3.3 Affiliation

In our prior work of PubMed Knowledge Graph [2], we have used the MapAffil 2016 dataset which contains PubMed authors' affiliations including cities and their geocodes to process the affiliations. All the affiliation strings were processed using MapAffil to identify and disambiguate the most specific places [12]. Same like our prior work about PubMed Knowledge Graph, we applied an up-to-date open-source library, Affiliation Parser36, to extract additional fine-grained affiliation fields from all affiliations, which included department, institution, email, ZIP code, location, and country. Our PubMed knowledge graph has processed data till early 2020. Since most of the COVID-19 articles have been published after March 2020, we use the PubMed Knowledge Graph to obtain affiliation information of authors who have already published articles in PubMed before March 2020. 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.

3.4 PubMed Knowledge Graph

We connect the CORD-19 dataset with the PubMed Knowledge Graph to build a strong knowledge base. Our lab has developed the PubMed Knowledge Graph covering 29 million PubMed articles using BioBert, disambiguated author names, integrated funding data through NIH ExPORTER, affiliation history and educational background of authors from ORCID, and fine-grained affiliation data

from MapAffil. By integrating the credible multi-source datasets, this PubMed knowledge graph contains connections among bio-entities, authors, articles, affiliations, and funding. We connect the PubMed Knowledge Graph with a drug discovery knowledge graph developed by Data2Disovery [3] to enrich the relationships of bio entities in the PubMed Knowledge Graph (which are the co-occurrence relationships) by using the biological relationships provided by this drug discovery knowledge graph. 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.

3.5 Clinical Trails

On Feb 25, 2020, the US started the first clinical trial to evaluate the safety and efficacy of remdesivir in hospitalized COVID-19 patients. Remdesivir is an antiviral treatment for humans with Ebola virus and has shown promise for treating Middle East Respiratory Syndrome (MERS) and severe acute respiratory syndrome (SARS), which are caused by virus from the same family of coronaviruses. To date, there are around 200 COVID-19 related clinical trials including drugs for HIV, malaria, experimental compounds working against viruses, and antibody-rich plasma from people who have recovered from COVID-19 [13]. We connected these clinical trials with our CORD-19 dataset through authors and institutions.

4 Result

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

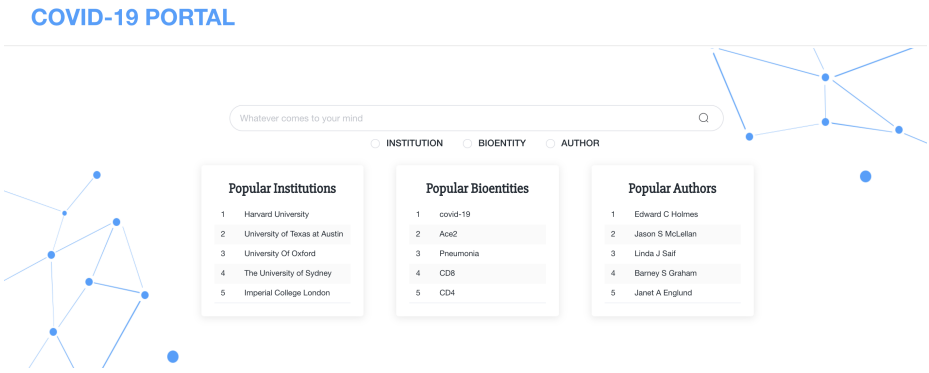

Fig. 1. Front page of the COVID-19 portal

4.1 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. User also can choose the simple search box to type their search terms. Before they start to type their search terms, a friendly reminder will pop out to reminder them to select one of the three categories first. Auto complete has been enabled when a user is typing an institution, an author, or a bio entity. The relevant auto complete will show up to facilitate this user to complete the search terms quickly and accurately.

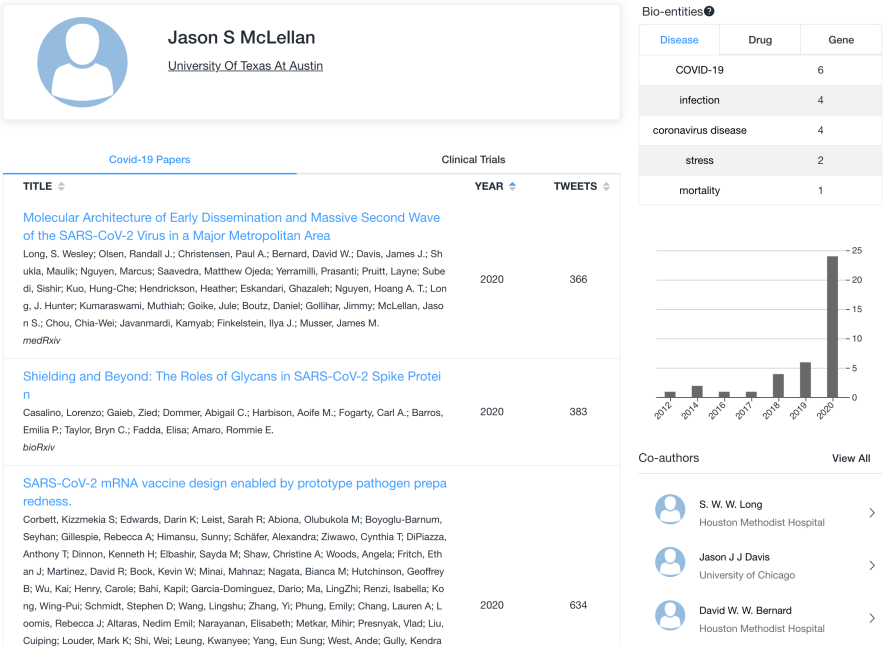

Fig. 2. Author profile with papers, bio-entities and coauthors

4.2 Profiling COVID-19 Scientists

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 public awareness of individuals working on COVID-19; 2) raise the scientific community's awareness of who is

working on what, with the goal of facilitating potential collaborations; and 3) show the international or inter-regional collaborations of scientific teams and inspire potential collaborations to fight against COVID-19.

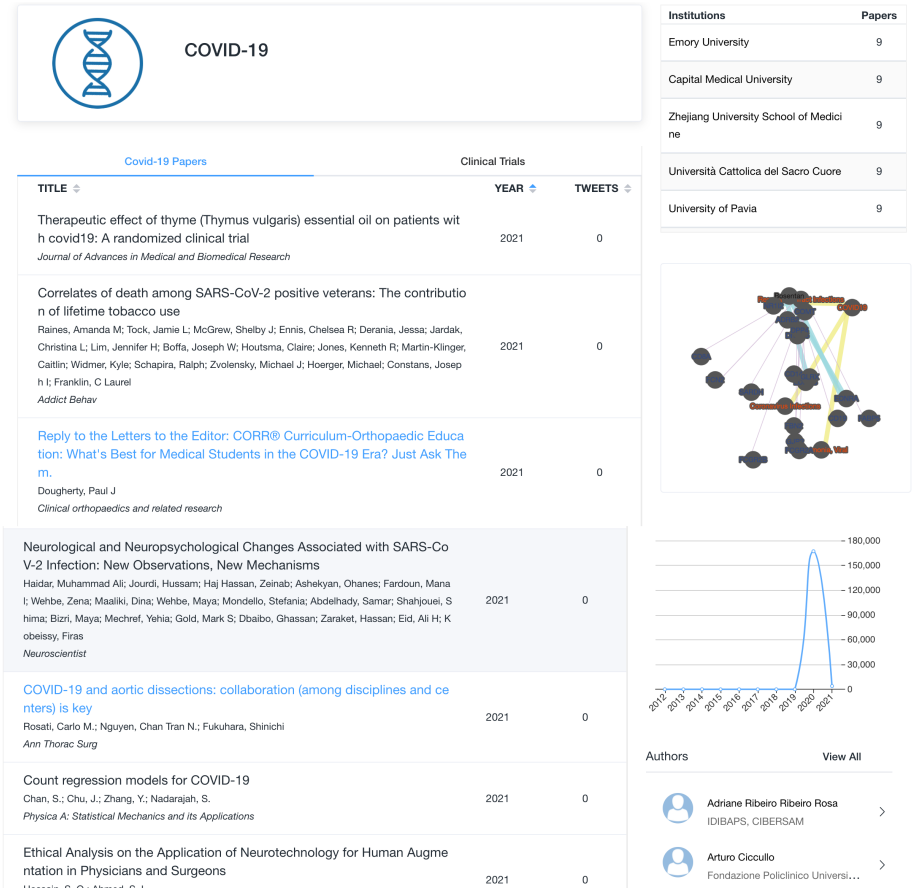

Fig. 3. Bioentity profile with papers, institutions, relationship graph, trends and coauthors

4.3 Profiling COVID-19 Bio Entities

For the bio entity profiling page (see Figure 3), the left side is the list of publications or clinical trials related to this bio entity ranked by article' publication year or the number of tweets. The right side is the list of institutions whose researchers published articles related to this bio entity, bio entity graph including related bio entities to current bio entity, the publication trend of this bio entity, and researchers who are working on this bio entity. Users can click papers,

institutions, and authors to go to their profile pages. For bio entity graph, we collaborate with Data2Discovery to extract bio entity relationships. Our portal includes 7 relationships and extracts top 20 related bio entities based on edge score provided by Data2Discovery.

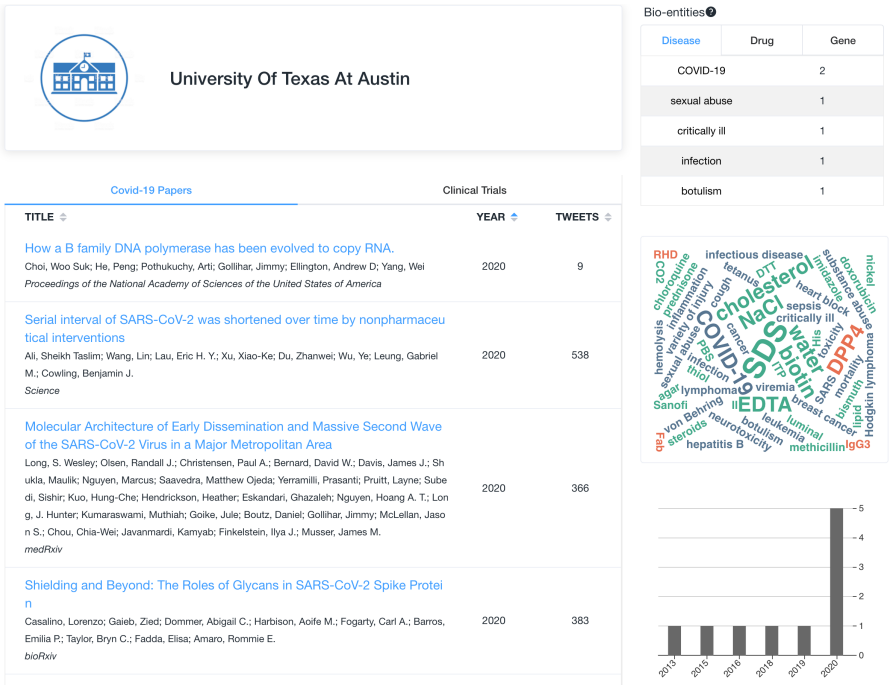

Fig. 4. Institution profile with papers, institutions and coauthors

4.4 Profiling COVID-19 Institution

For the institution profiling page (see Figure 4), the left side is the list of articles and clinical trials published by authors from this institution which can be ranked by publication year or the number of tweets. The right side is the list of bio entities which are mentioned in this list of articles and the bio-entities cloud which contains bio-entities from this list of publications. It also has the publication trend and a list of researchers from this institution. All authors, bio entities, and institutions on each page can be clicked and users will be directed to their corresponding profile pages.

5 Conclusion

The world has just experienced one of the biggest lock downs in human history due to the COVID-19 pandemic which has caused the tragic loss of human lives and the worst economic downturn since the Great Depression. 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.

References

1. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk,K., Kinney, R., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset.
2. 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.
3. 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.
4. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020 May;20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1.
5. COVID-19 Primer: https://covid19primer.com/dashboard.
6. COVID-19 Authors: https://covidauthors.org/search.
7. Wan, A., Zhan, Y., Tripathi, S., Yang, J., Sachdev, M., Paithankar, S., Duerksen, J., Chen, B., & Ding, Y. (2021). Building the COVID-19 portal by integrating literature, clinical trials, and knowledge graphs (poster). The ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL), September 27-30, 2021.
8. Wei, C., Allot, A., Leaman, R., Lu, Z. (2019). PubTator Central: Automated concept annotation for biomedical full text articles. Nucleic Acids Research, 47(1), W587-W593.
9. Aronson, A. R., Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of American Medical Informatics Association, 17, 229-236.
10. Bhatia, P., Celikkaya, B., Khalilia, M., Senthivel, S. (2019). Comprehend medical: a Named entity recognition and relationship extraction web service.
11. Lee, Jinhyuk and Yoon, Wonjin and Kim, Sungdong and Kim, Donghyeon and Kim, Sunkyu and So, Chan Ho and Kang, Jaewoo. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234-1240.
12. Torvik, V. I. (2015). MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. D-Lib Magazine: The Magazine of the Digital Library Forum. Vol. 21. No. 11-12. NIH Public Access.

13. Kupferschmidt, K., Cohen, J. (2020). Race to find COVID-19 treatments accelerates. Science, 367(6485), 1412-1413.