COVIDia: An Academic Knowledge Graph for COVID-19 Interdisciplinary Research

Cheng Deng[1], Jiaxin Ding[1], Luoyi Fu[1], Weinan Zhang[1], Jianghao Wang[2]
Xinbing Wang[1], Chenghu Zhou[1,2]
[1]Shanghai Jiao Tong University,
[2]Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences
davendw@sjtu.edu.cn
Corresponding authors: Jiaxin Ding and Luoyi Fu

ABSTRACT

The pandemic of COVID-19 has inspired extensive work across different research fields. However, extant literature and knowledge repositories pertaining to COVID-19 predominantly emphasize collecting papers on biology and medicine, neglecting interdisciplinary efforts. This lacuna impedes knowledge dissemination and collaborative research efforts across disciplines, which are essential for comprehensively addressing the multifaceted challenges posed by the pandemic. Studying interdisciplinary research requires effective paper category classification and efficient cross-domain knowledge extraction and integration. To this end, we introduce **COVIDia**, **COVID**-19 **i**nterdisciplinary **a**cademic knowledge graph, designed to serve as a nexus between COVID-19-related knowledge spanning different domains. The construction of the COVIDia is predicated upon disciplinary classification, knowledge entity linking, relation classification, and ontology management, all within the purview of interdisciplinary research. Leveraging the COVIDia, we develop various applications such as semantic search, geographic search, and a COVID-19 knowledge-enhanced question and answer. Furthermore, we provide benchmarks based on COVIDia including community detection, link predication, and COVID-19 related text corpus within the realm of interdisciplinary research. Finally, all the resources are publicly available. [1]

CCS CONCEPTS

• **Computing methodologies → Information extraction**; **Semantic networks**; *Ontology engineering*; *Reasoning about belief and knowledge.*

KEYWORDS

COVIDia, Academic Knowledge Graph, Information Extraction, Document Classification, Interdisciplinary Research.

ACM Reference Format:
Cheng Deng[1], Jiaxin Ding[1], Luoyi Fu[1], Weinan Zhang[1], Jianghao Wang[2], Xinbing Wang[1], Chenghu Zhou[1,2]. 2023. COVIDia: An Academic Knowledge Graph for COVID-19 Interdisciplinary Research . In *Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG '23).* ACM, Tokyo, Japan, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

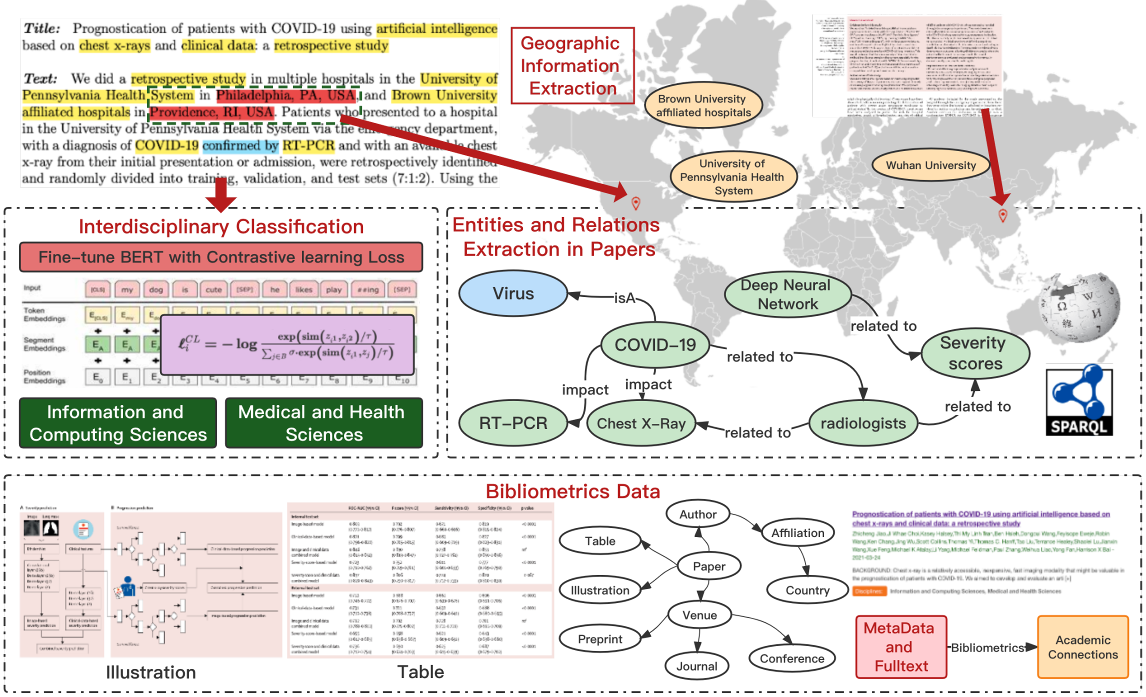
1 INTRODUCTION

The pandemic of COVID-19 has aroused the extensive attention of academic researchers worldwide, leading to innumerable lines of work radiating outward from COVID-19. Due to the profound and complex impact of COVID-19, the research works do not only include scientific discoveries in biology and medicine [20], but also require collaborations from other fields, such as computer science, sociology, mathematics, politics, etc. [10, 16, 42] to solve this problem with interdisciplinary insights [24]. Such boom in COVID-19 publications and the highly interdisciplinary collaborations even result in a "paperdemic" [13, 45], which makes qualified information and knowledge retrieval harder. Organizing interdisciplinary works and extracting knowledge is a critical problem to solve, which, however, has not been paid enough attention to.

Existing COVID-19 literature datasets focus on collecting papers on biology and medicine [7, 38, 41] without considering publications in other fields [9]. According to our statistics shown in Table 2, 48% out of 2.1 million publications on COVID-19 are not published in the venues of biology or medicine. Besides, even though part of interdisciplinary works are published in the biology and medicine venues, these works are not properly classified into all other disciplines they belong to [37], which makes it hard for researchers to draw interdisciplinary insights, and difficult for interdisciplinary researchers to find exact matches of their interests. Organizing interdisciplinary works is not simply putting together all papers related to COVID-19 but organizing works into the exact categories labeled by all related disciplines. Further, knowledge extraction and integration in interdisciplinary research are not well studied, which makes knowledge isolated in respective fields. For example, COVID-KG [42] extracts the bio-pharmaceutical and protein entities and establishes relations between entities from papers on COVID-19 in biomedical fields, while GAKG [11] mines the knowledge entities in the geoscience area. The ways of extracting knowledge entities in these different disciplines are different; the knowledge entities vary, and even the same words in different fields can have different meanings. Therefore, extracting knowledge from interdisciplinary works and integrating the corresponding knowledge extracted is challenging.

In the face of the above challenges, we propose frameworks to solve interdisciplinary paper classification and knowledge extraction. First, we propose a multi-label paper classification model with contrastive learning on different disciplines. Thereafter, we enhance the entity extraction model to interdisciplinary scenarios by aligning the entities to contents in Wikipedia to effectively extract entities with disciplinary contexts and efficiently adapt our model to open interdisciplinary research domains. We also propose an academic knowledge scheme graph, where knowledge entities on ontology layers can be related by sharing the same papers or

[1] https://github.com/davendw49/covidia



Title: Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study

Text: We did a retrospective study in multiple hospitals in the University of Pennsylvania Health System in Philadelphia, PA, USA, and Brown University affiliated hospitals in Providence, RI, USA. Patients who presented to a hospital in the University of Pennsylvania Health System via the emergency department, with a diagnosis of COVID-19 confirmed by RT-PCR and with an available chest x-ray from their initial presentation or admission, were retrospectively identified and randomly divided into training, validation, and test sets (7:1:2). Using the

Geographic Information Extraction
Brown University affiliated hospitals
University of Pennsylvania Health System
Wuhan University
Interdisciplinary Classification
Fine–tune BERT with Contrastive learning Loss
Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
Segment Embeddings
Position Embeddings
$\ell_i^{CL} = -\log \frac{\exp(\mathrm{sim}(z_{1,i}, z_2)/\tau)}{\sum_{j\in B} \sigma \cdot \exp(\mathrm{sim}(z_{1,i}, z_j)/\tau)}$
Information and Computing Sciences
Medical and Health Sciences
Entities and Relations Extraction in Papers
Virus
isA
COVID–19
Deep Neural Network
related to
Severity scores
related to
impact
impact
RT–PCR
Chest X–Ray
related to
radiologists
related to
SPARQL
Bibliometrics Data
Illustration
Table
Author
Affiliation
Table
Paper
Illustration
Country
Venue
Preprint
Journal
Conference
MetaData and Fulltext
Bibliometrics
Academic Connections
Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study
Discipline
Information and Computing Sciences, Medical and Health Sciences

Figure 1: COVIDia Overview: From Paper to Knowledge, we adopt information extraction, text classification, and bibliometric data management to organize the Academic Knowledge Graph of COVID-19

bibliometric entities on instance layers, and vise versa, for interdisciplinary knowledge integration.

In this work, we collect interdisciplinary papers on COVID-19, with the size of 2.1 million, and propose COVIDia, a **COVID**-19 **i**nterdisciplinary **a**cademic knowledge graph (KG), to address the problems of information retrieval and knowledge extraction of interdisciplinary researches on COVID-19. An overview of COVIDia with examples is demonstrated in Figure 1, the entire system and resources will be publicly accessible online, and our contributions can be summarized as follows:

1. In this paper, we propose an ongoing interdisciplinary academic knowledge graph on COVID-19, **COVIDia**, which summarizes bibliometrics data, domain glossaries, illustrations, tables, and spatiotemporal information of papers and relations between them. To our knowledge, COVIDia is the largest knowledge graph and academic research literature platform for COVID-19.
2. This paper achieves interdisciplinary paper classification by introducing contrastive learning over interdisciplinary categories. We introduce an entity extraction model based on learning to rank and a BERT-based relation extraction model using segment embedding. We deploy this model and extract knowledge entities in the COVID-19 articles and relations between these entities.
3. In COVIDia, we put forward a new academic knowledge graph scheme, adding disciplinary categorization to knowledge entities, which facilitates researchers on information retrieval and mining.
4. Based on COVIDia, we provide a dump dataset of the KG, *four* benchmarks, including tasks of network science and natural language processing, as well as *three* applications for data mining and information retrieval.

2 RELATED WORK

We briefly review the related work of COVIDia, including literature classification, information extraction, and COVID-19 related KG.

Literature Classification. Since the outbreak of COVID-19, researchers have employed natural language processing models and developed learning methods to understand the pandemic-related text material [8]. [17] provides an analysis of several multi-label document classification models on the LitCOVID dataset, and pretrained language models [35] outperform others. [26] use label correlation to estimate the similarity between papers. [37, 40] classify and evaluate interdisciplinary papers by designing catalogs and indicators.

Information Extraction in Literature. When building a KG, mining entities and relations are the main challenges. In the task of entity extraction, BERT-based [12] and BiLSTM-CRF [22] are mainstream models, which have been deployed in domain-specific scenarios like material science [43] and geoscience [11]. AutoPhrase [33] sheds light on unsupervised phase tagging. For relation extraction, the works of mining relations between entities extracted from literature are rare and mainly in biomedical articles [1, 4, 7]. Meanwhile, the means like OpenRE [18] share the idea of distant supervision on close-domain relation extraction, while OpenIE [3] gives the open-domain task a brand new paradigm.

COVID-19 related Knowledge Graph. In COVID-19 scenarios, dimensions.ai [27] provides a subset of Digital Science via a set of keyword queries, while CORD-19 [41] provides a machine-readable research dataset. COVID-KG [42], COVID-19 KG RDF database [6] and KG-COVID-19 [30] has collected the COVID-19

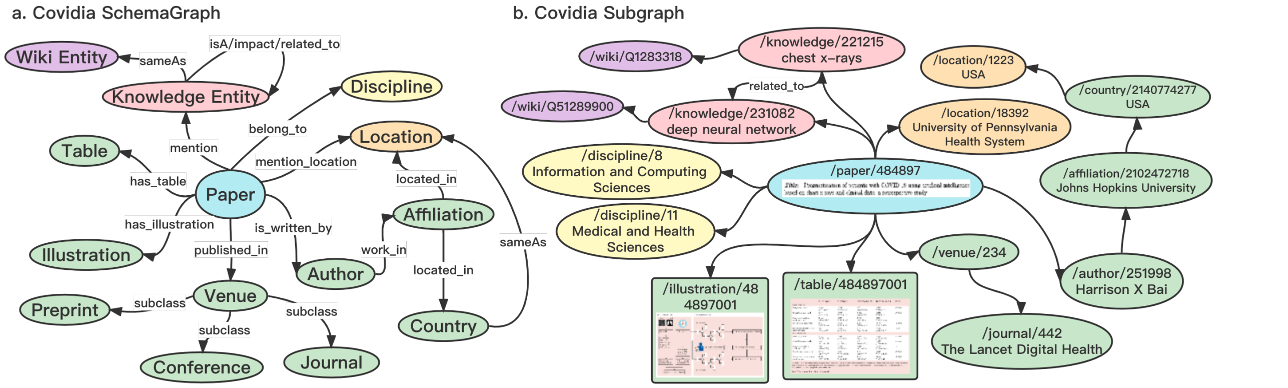


Figure 2: COVIDia Schema, Figure a. is the schema-graph of COVIDia and Figure b. is an example and a subgraph of COVIDia.

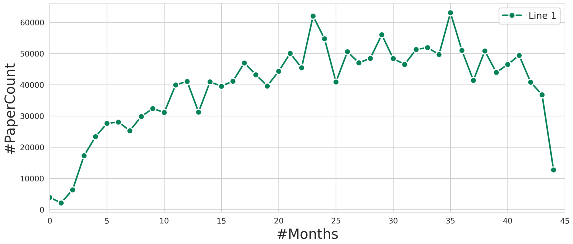


Figure 3: Monthly changes in the amount of data added, according to COVIDia.

Table 1: Statistics Concepts and Relations in COVIDia.

Concepts	Count	Concepts	Count	Relations	Count	Relations	Count
paper	2,102,872	topic	17,334	is_cited_by	5,179,621	work_in	2,370,233
author	3,180,665	discipline	22	is_written_by	8,911,921	is_located_in	314,158
organization	19,571	papertable	1,247,876	is_published_in	2,102,872	has_papertable	1,247,876
journal	69,364	illustration	1,811,651	in_the_topic_of	2,041,288	has_illustration	1,811,651
conference	9,205	knowledge	1,569,310	belongs_to	2,903,771	related_to	6,129,291
preprint	15	location	169,451	mention_knowledge	3,228,392	rdfs:subClassOf	45,231
venue	78,584	geohash	169,451	mention_location	1,250,738	owl:sameAs	2,119,181
Total Entities		**10,366,787**		**Total Relations**		**39,656,224**	

related literature metadata and the key terms related to the papers. Cause-and-effect KG on COVID-19 pathophysiology is proposed by [14]. A framework that can integrate heterogeneous biomedical data to produce KGs is developed for COVID-19 [30].

3 COVIDIA OVERVIEW

3.1 Motivation of Constructing COVIDia

Although WHO has declared "with great hope" an end to COVID-19 as a public health emergency [2] and we also observe a drop of COVID-19 related publications, shown in Figure 3, the disease is still a global threat and research works are still going on. In retrospect, the evolution of COVID-19 research works from the beginning to the "end" of the era of global public pandemic provides us an unprecedented opportunity to study the interdisciplinary knowledge evolution on a brand new topic intensively in a short period of time. A deeper understanding of how we can better collaborate among different fields or disciplines to solve such a complicated crisis can prepare us more on the future unexpected potential emergency.

We notice that many collections of COVID-19-related research works released in the early pandemic are no longer well-maintained. To systematically study COVID-19 works, it is crucial to maintain collections of academic researches across all fields during the entire pandemic period.

Therefore, we are motivated to build COVIDia . It is an academic knowledge graph that collects cross-disciplinary research during the pandemic period. We have gathered literature and related academic elements from all disciplines over the past three years of the pandemic, as well as the relationships between knowledge entities mentioned within the articles.

3.2 The COVIDia

We first present the overview of COVIDia. There are **13** concepts connected by **15** relations in the COVIDia and **34** data properties. The distribution of the disciplines, are shown in Table 2. We can see that about half of the publications on COVID-19 are not published in the venues of biology or medicine science. To obtain fine-grained knowledge in COVID-19 papers, we design the concept *covidia:knowledge* representing the knowledge entities extracted from the papers. Meanwhile, COVIDia use relation *covidiar:mention_knowledge* to connect the knowledge entities and papers, design relations *covidiar:is_A*, *covidiar:increase*, *covidiar:decrease* and *covidiar:related_to*, along with *covidiar:related_to* is also acting as the abstract axioms of *covidiar:increase*, *covidiar:decrease* and *covidiar:is_A*.

In COVIDia, **we also adopt *owl:sameAs* axioms linking knowledge entities to the DBpedia [5] entities,** so as to link COVIDia to the outer KGs. Furthermore, *geohash* is used to represented the points of interest (POI). To sum up, the base namespace (Graph IRI) for COVIDia is https://covidia.acemap.info, shared among all the concepts and relations. The schema-graph of the COVIDia is shown in Figure 2.

The statistics of COVIDia concepts and relations is shown in Table 1. The knowledge graph consists of **6,097,866** instances and **24,705,417** links. To facilitate scholars in data mining, knowledge engineering, and information retrieval to browse and access COVIDia data efficiently, we provide a SPARQL endpoint. and use SNORQL [2] to encapsulate it. Moreover, programmers can access the data through tools like sparqlwrapper with setting IRI links.

[2] https://github.com/kurtjx/SNORQL

Table 2: Disciplines Distribution in Covidia.

Discipline	Count	Discipline	Count
Mathematical Sciences	206,651	Medical and Health Sciences	862,080
Physical Sciences	103,523	Built Environment and Design	29,997
Chemical Sciences	119,490	Education	92,892
Earth Sciences	17,198	Economics	56,758
Environmental Sciences	85,103	Commerce, Management, Tourism and Services	75,453
Biological Sciences	304,991	Studies in Human Society	195,012
Agricultural and Veterinary Sciences	33,647	Psychology and Cognitive Sciences	118,497
Information and Computing Sciences	145,152	Law and Legal Studies	26,991
Engineering	234,416	Studies in Creative Arts and Writing	11,311
Technology	21,642	Language, Communication and Culture	78,571
Medical and Health Sciences	862,080	History and Archaeology	48,190
Built Environment and Design	29,997	Philosophy and Religious Studies	36,206

Table 3: Statistics of the COVIDia Data Sources.

	Paper	Author	Organization	Venue
Acemap	354,940	725,525	10,161	14,438
CORD-19	970,835	1,445,982	81,029	131,002
Digital Science	1,790,530	1,911,861	1,661,993	42,198
Preprint Website	229,098	198,198	2,326	15
After Fusion	2,102,872	3,180,665	19,571	78,584

4 COVIDIA CONSTRUCTION

In this section, we detail processes of building COVIDia (as shown in Figure 4, [29] share the figure of interdisciplinary research). We first obtain data integrated from different paper sources and perform interdisciplinary paper classification to classify the papers. Meanwhile, we tag the specialized terms mentioned in the articles. We perform open-domain information extraction for each COVID-19-related paper, align the entity to the glossary, and classify their relations.

4.1 Bibliometrics Data Collection and Fusion

We fuse the data by integrating papers from AceMap [36], CORD-19 [41], Digital Science [27], and 15 preprint sites that were flagged as being related to COVID-19, normalizing institutions, and naming scholars. The details of papers, authors, organizations and venues for each data source are listed in Table 3.

We extract the papers marked as *"2019 20 coronavirus outbreak"*, *"severe acute respiratory syndrome coronavirus 2"* and *"Coronavirus disease 2019"* in Acemap, collect the papers published by CORD-19 since January 2020, and collected the COVID-19 papers collected by Digital Science, using the query shared [27], with the type of *"article"*, *"proceeding"* and *"preprint"*. At last, we collect papers that study COVID-19 from **15** mainstream preprint websites. Each data source focuses on a different application scenario. We adopt scholars' name disambiguation and organization normalization algorithms [36] to finish the information fusion. We build a BERT-based named entity recognition model and use SpaCy [19] tools to extract the geographical and political locations. After strict normalization rules of the location entities, we deploy a GeoCoder to get the location coordinates. Finally, we use the Geohash algorithm to store coordinates to present the distribution of COVID-19 papers on a world map online.

All above, COVIDia gathers all papers, scholars, institutions, and related bibliometric information about COVID-19 with papers' illustrations, tables, knowledge entity, and geographic locations extracted from the text.

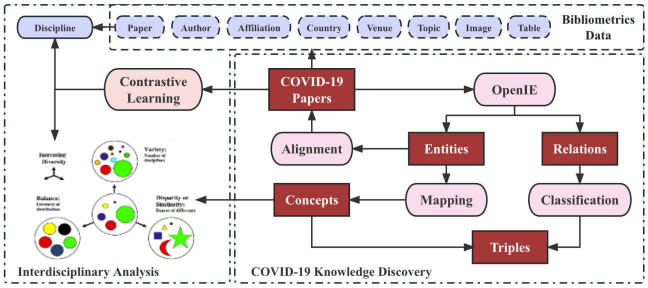


Figure 4: Construction Pipeline of COVIDia.

4.2 Interdisciplinary Paper Classification

Pre-trained language models based document classification model are the most widely used [17]. Similar to [35], we develop an interdisciplinary classification model that combines the embeddings generated by SciBERT and contrastive learning loss with **220,330** interdisciplinary annotated labels from Acemap.

To perform single-task fine-tune BERT, we first choose the *BCE Loss* with logits viewing interdisciplinary paper classification as multiple binary classification,

$$\ell_k^{BCE} = y_i \log(x_i) + (1 - y_i) \log(1 - x_i), \tag{1}$$

where k is denoted as the index of the batch, for each sample with index i, x_i is the predicted label, y_i is the ground true label. Meanwhile, intuitively two articles with totally different labels should be placed with a longer distance in the latent space. Similar to [39, 46], COVIDia adopts a comparative learning loss function *InfoNCE Loss*, for each document z we generate a pair of augmentations for each sample in a batch B (with a size $|B|$), to reduce the loss as:

$$\ell_i^{CL} = -\log \frac{\exp\left(sim\left(z_{i^1}, z_{i^2}\right)/\tau\right)}{\sum_{j \in B} \sigma \cdot \exp\left(sim\left(z_{i^1}, z_j\right)/\tau\right)}, \tag{2}$$

where z_{i^1}, z_{i^2} is a positive pair, σ is an indicator function and τ denotes the temperature parameter setting as 0.5. The final loss function is:

$$Loss = \ell_k^{BCE} + \ell_k^{CL}. \tag{3}$$

We employ this model as an encoder to generate all the embeddings of the paper in COVIDia.

4.3 Knowledge Entity Extraction

Distribution of discipline glossary can provide potential connections between papers that have neither citation relations nor the same authors. Now we introduce the pipeline to extract knowledge entities and the referred locations from papers text.

For 22 disciplines, we have collected high-quality wiki entities, corresponding discipline glossary sets and discipline knowledge graphs as a set of disciplines knowledge. We ensure that each discipline has a high-quality discipline glossary, with disciplines knowledge graphs as replenishment. Meanwhile, for the sake of disambiguation, we invite experts from various disciplines to separate entities and claim them as the concept *COVIDia:knowledge*, so that they can be linked to the entities in the original knowledge graphs through the assertion of the *#sameAs*.

With the above disciplines knowledge sets, discipline knowledge graphs and wiki entities, we need to match the sentences in the papers semantically. Once the sentences in the text are semantically close to or directly refer to the discipline glossary sets and discipline knowledge entities, we can claim the relation of *mention_knowledge*. Therefore, referring to explicit semantic analysis (ESA) [15], we first convert the terms that may be associated in an abstract through the vector transformation of TF-IDF [31]. In this process, we regard all the glossary and nodes in knowledge graphs in the table as entities, and their descriptions as documents D, view papers' abstracts as a query q to find the words in the text. Thus, we can get the candidate entities E.

$$E = Q(q, D), \quad E = \{e_i\}, i \geq 0. \tag{4}$$

Second, we take each entity's TF-IDF score, length, complexity, and letters' amount as the feature vector, consulting the LambdaRank [28], one of the learning to rank algorithms, we try to learn the function where given a text q, return with n-dims entities E, with related n-dims scores S, like $f(q, E) = S, \quad E = e_i, S = s_i, i \geq 0,$. Then we train a two-layer neural network, combining the feature vectors of entities and a loss function towards a pair as equation 6 to ensure that factor NDCG [23] can finally reach a certain level.

$$Loss_{ij} = \log\left\{1 + \exp\left(-\sigma\left(s_i - s_j\right)\right)\right\} \cdot \left|\Delta NDCG_{ij}\right|, \tag{5}$$

where σ is a parameter shaping the *sigmoid* function. The binary label we used to calculate the NDCG is annotated by experts from various disciplines, according to every pair of paper and candidate entities generated by the ESA step.

Finally, by setting the threshold, a considerable result is obtained. In this process, we sacrificed the recall rate in order to ensure accuracy. Overall, the precision of our model on the benchmark we set is **0.914**, and the recall rate is **0.391**.

4.4 Relation Classification Between Knowledge Entities

In addition to bibliometric data, COVIDia also mines the relations between knowledge entities mentioned in the papers to construct a knowledge graph for each paper. By recognizing the relations between knowledge entities of the papers, research can do reasoning over knowledge entities. Referring to the definitions of knowledge engineering and common sense knowledge graph, we define three relations *is_A*, *increase*, *decrease*, and *related_to*. We first extract triples through the general open-domain triple extraction tool *OpenIE* [3]. For each paper, we extracted an average of **53** triples. For each paper p we have

$$OpenIE(p) = (h_o, r, t_o), \tag{6}$$

where h_o and t_o denotes the origin entities, and r indicates the relations extracted by the *OpenIE*. Then, we align the obtained head entity and tail entity with the entity obtained in the previous section by adopting the exact match rule to ensure accuracy to align and get the normalized head entity and tail entity h and t. After that, we obtain persistent annotation data by deploying the Human-in-the-Loop system. We set the obtained annotations quadruple *(sentence, head entity, relation, tail entity, label)* as (s, g, r, t, l), where l is *covidiar:is_A*, *covidiar:has_impact*, *covidiar:is_A*, and *unknown*.

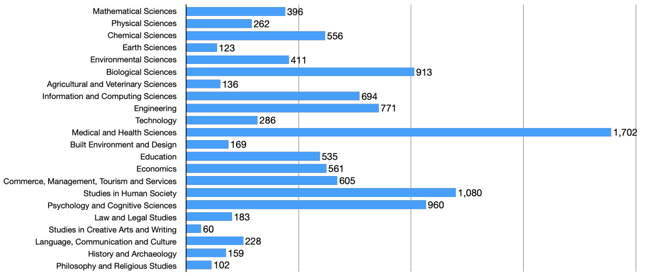


Figure 5: The discipline distribution of the training set.

Based on BERT, we design a linear layer as the task-specific layer for relation classification. The input to BERT is a single sentence with annotated entities and relations. Then we use the sum of three vectors to encode the input, of which token embedding is the WordPiece word vector, and segment embedding is specially designed, (embedding value 1 if the token is one of the elements in the triple, and the rest is 0), and a trainable position embedding. Finally, the probability is obtained through the $Softmax$ layer, and the output probability is as follows,

$$p(l \mid \mathbf{h}) = Softmax(\mathbf{Wh}), \tag{7}$$

where $\mathbf{h}$ stands for the output of the encoder and $\mathbf{W}$ is the parameter of the linear layer. Furthermore, the model parameters are optimized by maximizing the log-probability of the correct label. Finally, we process the model over the entire COVID-19 papers.

5 BENCHMARKS AND DATASETS

This section will evaluate and demonstrate the methods we use to construct the COVIDia. We make benchmarks for interdisciplinary paper classification, knowledge entity extraction, and open domain relation classification on our dataset and compare several baselines. In interdisciplinary paper classification, we choose AUC as the measure, use precision and recall in the experiment for phrase extraction and open domain relationship classification, and obtain the final result by delineating a reasonable threshold.

5.1 Interdisciplinary Papers Classification

5.1.1 Dataset. We also constructed a mapping of COVID-19 articles and discipline labels based on the COVIDia. Each article has at least two to four discipline labels. Our discipline labels are collected from digital science. The entire dataset contains 40,000 articles, including the articles' titles, abstracts, and IDs. For this dataset, the distribution of disciplines is as Figure 5.

5.1.2 Evaluation. We assessed our method and some other baselines on the task of interdisciplinary paper classification. We compare LAHA [21], **SciBERT+FT**, and **SciBERT+CL** where SciBERT+FT is a model finetuning the SciBERT with BCE Loss on language model downstream interdisciplinary paper classification task; SciBERT+CL is the model using contrastive learning mechanism.

As shown in the Table 5, the model trained with contrastive learning loss has the highest AUC, although the precision and ranking metric NDCG is comparable. In order to enable the embedding of articles to contain classification information of different disciplines, we use the contrastive learning-based model.

Table 4: Statistics of COVIDia Social Science Benchmarks.

Networks	Concepts	Size	Volume	Max Degree	Avg Degree	α	p_value	x_min
Coauthor Network	author	303,995	4,853,932	1,286	15.96	1.436	0.613	246
Citation Network	paper	398,920	5,218,680	15,980	13.08	1.503	0.107	184
Author-Paper Network	author	2,636,703	5,444,318	2,770	2.06	1.511	0.912	121
(Author-writes-Paper)	paper	967,070		5,368	5.63	1.32	0.748	104
Paper-Author Network	author	981,419	17,418,387	2,701	17.75	1.377	0.193	401
(Paper-inspires-Author)	paper	247,338		99,836	70.42	1.569	0.737	596

[1] α, p, x_{min} is used to measure the power-law distribution characters.

Table 5: Experimental results of interdisciplinary paper classification.

	Pre.@3	Pre.@5	NDCG@3	NDCG@5	AUC
LAHA	88.34	**77.78**	91.31	89.30	95.12
SciBERT+FT	87.99	76.58	91.18	88.48	95.24
SciBERT+CL	**89.20**	77.61	**92.17**	**90.32**	**96.33**

Table 6: Experimental results of Phrase Tagging

Methods	Precision	Recall
ESA+BERT+L2R (1st Loop)	0.859	0.292
ESA+BERT+L2R (2nd Loop)	0.872	0.331
ESA+Word2Vec+L2R (1st Loop)	0.841	0.319
ESA+Word2Vec+L2R (2nd Loop)	0.854	0.347
ESA+TF-IDF+L2R (1st Loop)	0.881	0.332
ESA+TF-IDF+L2R (2nd Loop)	**0.914**	**0.391**

5.2 Knowledge Entities Linking

5.2.1 Dataset. By deploying an annotation system for Human-In-The-Loop machine learning, we allow experts in all disciplines to score our predicted phrases and build a ranking dataset to find the top entities in tagging phrases. The dataset has collected more than 1000 articles with 50,000 phrase annotations.

5.2.2 Evaluation. For the phrase tagging task, we compare **ESA+L2R** model using TF-IDF, using BERT and Word2Vec as the feature constructing on our dataset. Since we have manual annotations to help improve the model's performance in cycle to guarantee the precision, we sacrifice the recall so that our model can tag out more accurate phrases. Meanwhile, we choose the best threshold on the test set and test the selection over the benchmark. The results are shown in Table 6, showing that the mechanism using TF-IDF in the second loop performs better than the other models.

5.3 Relation Classification

5.3.1 Dataset. Relation classification is the task of identifying the semantic relation between two entities in the text. With the help of the annotation system, we gather more than two thousand *(triple, sentence, label)* records. For the open domain, we have designed two baselines for comparison to classify the relation in the open domain into three categories, including *is_A*, *subclass_of* and a superclass *related_to* to generalize the rest.

Table 7: Experimental results of Open-domain Relation Classification.

	Precision	Recall
SciBERT+FT	76.53	71.18
SciBERT+SEG	**79.09**	**75.21**

5.3.2 Evaluation. We set 400 of them as the benchmark to evaluate the models. We use precision and recall to evaluate these two models, and the results are in Table 7, indicating that with re-designed segment embedding, the pre-train model can perform better.

5.4 Network Science Dataset

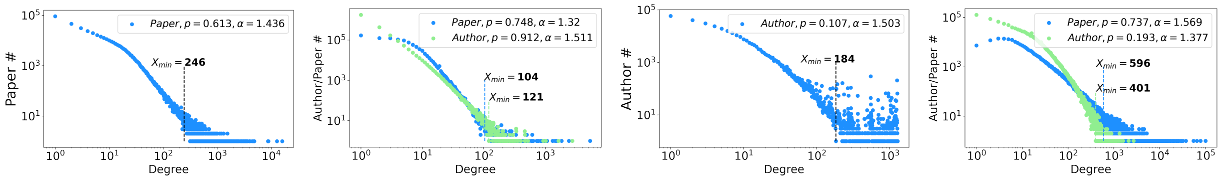
COVIDia provides four social networks to network science. There are two homogeneous networks, an author cooperation network and a citation network, which can be used for author and paper classification and COVID-19 community detection. Two paper-author networks are also obtained, including a bipartite graph network between articles written by authors and a bipartite graph network between authors and papers cited by their papers, which can be used for reference recommendations. The relevant data statistics and the corresponding distribution can be found in Table 4 and Figure 6.

6 COVIDIA APPLICATION

COVIDia has been applied to three COVID-19 academic scenarios, including but not limited to retrieval of COVID-19-related papers via POI and semantic search over COVIDia RDF dataset. Meanwhile, the COVIDia can also be applied to network data mining as a heterogeneous network.

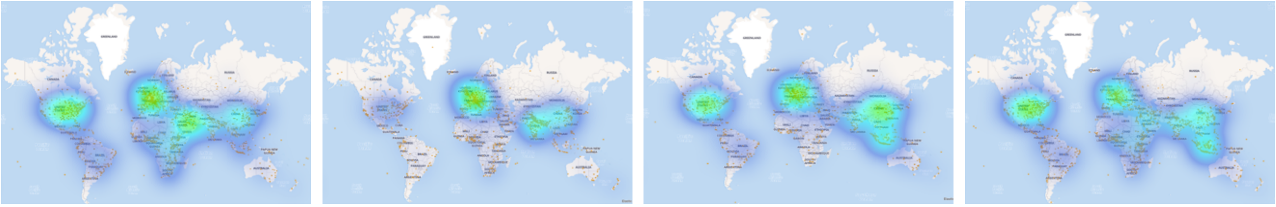
6.1 Geographic Search over COVIDia

We provide the geographic locations mentioned in each paper. By visualizing the results, as shown in Figure 7, we can see that the distribution of searched papers varies for different keywords. When we set the keyword to *"women"*, we can find that both Eurasia and North America are scorching areas. However, when we look for papers about *"lockdown"*, the difference between Eurasia and the Americas is apparent, which can be explained by different quarantine policies in different regions. Coincidentally, the discipline distribution of *"computer science"* tends to be slightly less than that of *"psychology"*.



(a) Citation Network (b) Author-Paper Network (c) Coauthor Network (d) Paper-Author Network

Figure 6: Degree distribution on Network Science Benchmarks.



(a) (b) (c) (d)

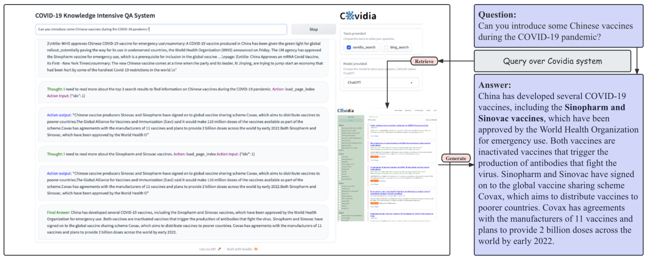
Figure 7: Papers distribution on earth regarding different topics. (a) Keywords: "Woman", (b) Keywords: "Lockdown", (c) Discipline: "Information and Computing Sciences", (d) Discipline: "Psychology and Cognitive Sciences".

Based on geography information in COVIDia, we provide a geographic search over COVIDia. First, researchers can input any phrases to query in COVIDia, including papers' titles, abstracts, disciplines, and knowledge entities the COVID-19 papers mentioned. Researchers enter keywords in the input box, e.g., "Vaccine in China", and the related papers would be shown on the map. Besides, for better user interaction, users can drag the map, zoom in, and zoom out the map to search for papers within the bounding box with specified keywords.

6.2 Semantic Search over COVIDia

Using COVIDia, researchers can learn more about the relationship in COVID-19 academia. COVIDia perform several examples, including one-hop queries, such as returning articles that mention a particular knowledge point; two-hop queries, such as retrieving illustrations in a particular area and retrieving places that an affiliation, usually studies as an example of a three-hop query. These queries can be used in scientific research and scholarly communication about COVID-19. Moreover, these queries are generally unanswerable by existing search engines and differ from the current academic platform.

6.3 COVID-19 knowledge intensive QA



COVID-19 Knowledge Intensive QA System

Can you introduce some Chinese vaccines during the COVID-19 pandemic?

2\ntitle: WHO approves Chinese COVID-19 vaccine for emergency use\nsummary: A COVID-19 vaccine produced in China has been given the green light for global rollout, potentially paving the way for its use in underserved countries, the World Health Organization (WHO) announced on Friday. The UN agency has approved the Sinopharm vaccine for emergency use, which is a prerequisite for inclusion in the global vaccine ...\npage: 3\ntitle: China Approves an mRNA Covid Vaccine, Its First - New York Times\nsummary: The new Chinese vaccine comes at a time when the party and its leader, Xi Jinping, are trying to jump-start an economy that had been hurt by some of the harshest Covid-19 restrictions in the world.\n"

Thought: I need to read more about the top 3 search results to find information on Chinese vaccines during the COVID-19 pandemic. Action: load_page_index Action Input: {"idx": 1}

Action output: "Chinese vaccine producers Sinovac and Sinopharm have signed on to global vaccine sharing scheme Covax, which aims to distribute vaccines to poorer countries.The Global Alliance for Vaccines and Immunization (Gavi) said it would make 110 million doses of the vaccines available as part of the scheme.Covax has agreements with the manufacturers of 11 vaccines and plans to provide 2 billion doses across the world by early 2022.Both Sinopharm and Sinovac, which have been approved by the World Health O"

Thought: I need to read more about the Sinopharm and Sinovac vaccines. Action: load_page_index Action Input: {"idx": 1}

Action output: "Chinese vaccine producers Sinovac and Sinopharm have signed on to global vaccine sharing scheme Covax, which aims to distribute vaccines to poorer countries.The Global Alliance for Vaccines and Immunization (Gavi) said it would make 110 million doses of the vaccines available as part of the scheme.Covax has agreements with the manufacturers of 11 vaccines and plans to provide 2 billion doses across the world by early 2022.Both Sinopharm and Sinovac, which have been approved by the World Health O"

Final Answer: China has developed several COVID-19 vaccines, including the Sinopharm and Sinovac vaccines, which have been approved by the World Health Organization for emergency use. Both vaccines are inactivated vaccines that trigger the production of antibodies that fight the virus. Sinopharm and Sinovac have signed on to the global vaccine sharing scheme Covax, which aims to distribute vaccines to poorer countries. Covax has agreements with the manufacturers of 11 vaccines and plans to provide 2 billion doses across the world by early 2022.

Tools provided
Choose the tools to solve your question.
☑ covidia_search ☐ bing_search

Model provided
Choose the model to solve your question, Default means ChatGPT.
ChatGPT

Question:
Can you introduce some Chinese vaccines during the COVID-19 pandemic?

Query over Covidia system

Answer:
China has developed several COVID-19 vaccines, including the **Sinopharm and Sinovac vaccines**, which have been approved by the World Health Organization for emergency use. Both vaccines are inactivated vaccines that trigger the production of antibodies that fight the virus. Sinopharm and Sinovac have signed on to the global vaccine sharing scheme Covax, which aims to distribute vaccines to poorer countries. Covax has agreements with the manufacturers of 11 vaccines and plans to provide 2 billion doses across the world by early 2022.

Figure 8: COVID-19 knowledge intensive QA System.

With the development and wide use of large models in academia and industry, many chatbots based on generated large language models have emerged. However, these chatbots cannot guarantee the correctness of the generated content. To this end, retrieval-augmented language models and combining tools and language models appear. During the outbreak of the pandemic, much misinformation has appeared online. If a language model is trained based on web-crawled data, it may capture many incorrect information.

In this paper, we have built a retrieval-enhanced QA system based on COVIDia using langchain and bmtool. At the same time, we have organized a batch of supervised data related to COVID-19 for instruction tuning on language models. Below, we present examples of using our system for pandemic-related question retrieval.

7 DISCUSSION ON COVIDIA

In line with the common data publication best practices and FAIR principles for scientific data management, in this session, we will discuss the availability, reusability, quality and related works of COVIDia. And finally share our point of views on COVIDia potential impact and direction in the future.

7.1 Availability, Reusability and Quality

Availability. Based on COVIDia, we put forward serveral resources, including a COVID-19 interdiciplainary knowledge graph dataset, 4 benchmark collections for network science, natural language processing and data minig, as well as 3 COVID-19-related applications.

In addition, we have a sustainability plan specified for long-term maintenance of COVIDia under the MIT License. The URI of each instance from COVIDia is reachable, to meet the needs of Linked Open Data.

Reusability. During the early stages of development, the platform functioned primarily as a data provider for various departments engaged in data analysis, encompassing publications published in the European Journal of Internal Medicine (EJIM). Recently, the World

Table 8: Comparison with related work on COVID-19 Collections.

Collections	Data source	Maintenance	Application	Location	PaperNum	Corpus	Discipline
Covidia	**Multiple**	Long term	**System, Data**	**GPE**	**2,102,872**	**full-text**	**Interdiscipline**
CORD-19	Semantic Scholar	2020.1-2022.6	Dataset	N/A	970,835	abstract	Biomedicine
LitCOVID	PubMed	Long term	**System, Data**	Country	349,546	abstract	Biomedicine
COVID-DP	EuropePMC	Long term	System	N/A	959,926	abstract	Biomedicine
Dimensions	Digital Science	Long term	System	Country	1,790,530	abstract	**Interdiscipline**

Health Organization (WHO) declared that COVID-19 no longer constitutes a "global health emergency." Nevertheless, the ongoing emergence of new cases and fatalities underscores the sustained relevance of COVID-19 as a research subject. As a valuable resource, will persist in offering its services to the scholarly community, facilitating further investigation and research. In the foreseeable future, the integration of large language models, data mining techniques, and semantic networks will gain increasing attention. possesses the capacity to bolster the correctness of generated models with responding content about COVID-19. Comprehensive instructions delineating the pertinent details are documented on Github.

Quality Evaluation. The construction algorithms of COVIDia have been developed to ensure objective *precision* on benchmarks while sacrificing *recall*. We also provide a user feedback mechanism to refine COVIDia, and will continuously update COVIDia auto-construction module in the future. The source code and experiment information are on Github.

7.2 Comparison with related work

We briefly review the related work of COVIDia, mainly on academic collections about COVID-19. Dimensions.ai [27] provides a subset of Digital Science via a set of keyword queries, while CORD-19 [41] provides a machine-readable research dataset. LitCOVID [7] share the COVID-19 papers of PubMed while COVID-19 Data Portal, (COVID-DP in brief), share the COVID-19 papers of EuropePMC. Based on CORD-19, COVID-KG [42], CKG [44], COVID-19 KG RDF database [6] and KG-COVID-19 [30] has collected the COVID-19 related literature metadata and some of the key terms related to the papers. COVID-on-the-Web [25] linked CORD-19 corpus and DBpedia. IMGT-KG [32] collect only immunogenetics data from various data sources. Moreover, a cause-and-effect KG on COVID-19 pathophysiology is proposed by [14], while some framework that can integrate heterogeneous biomedical data to produce KGs was developed for COVID-19 [30, 34]. And, the detailed statistics are shown in Table 8.

7.3 Potential Impact and Future works

Potential Impact. As mentioned above, COVIDia is the largest and most actively maintained COVID-19-related dataset and platform currently available, while also being the sole repository with a well-defined interdisciplinary classification. By following the principles of linked data and aligning with DBpedia, COVIDia enables entity alignment, which allows it to establish connections with other linked data and facilitate external SPARQL queries. Moreover, the text corpus provided by COVIDia supplies generative language models with structured knowledge regarding the COVID-19. Consequently, when utilizing these large models to generate content related to the pandemic, we can have more reliable information.

Future works and Directions. As we will maintain and update the COVIDia in the future sustainably. Based on COVIDia, future works and directions from our perspective are shared as follows:

1. Evolving Academic Community Detection in COVID-19. In addition to sharing the benchmarks that the community can use to evaluate existing models and algorithms, we also hope to discover how researchers cooperate after the outbreak and which scholars have played a vital role in the co-author network like bridges connecting groups from different disciplines.
2. COVID-19-related Misinfo Detection. With COVIDia, we can facilitate the generative language models and make the content generated more reliable, which can greatly shorten the time for scholars to read literature, improve the efficiency of surveys, and policymakers can more efficiently review and understand the situation of the emergency and obtain methodologies from them.
3. Interdisciplainery research. During the pandemic, the growth of interdisciplinary is a very special phenomenon. As a topic of COVID-19, scholars from different disciplines will participate in it. Therefore, with the help of COVIDia, the cooperation of researchers can have a better interdisciplinary research experience.

8 CONCLUSION

In this paper, we propose a novel COVID-19 interdisciplinary academic knowledge graph, **COVIDia**, which extracts knowledge from all COVID-19-related research papers published in the major venues across different disciplines. The framework can not only benefit the researchers on COVID-19 but also be leveraged to study potential future pandemics. The techniques we propose to generate interdisciplinary knowledge graphs are not limited to applications on COVID-19 but can also be applied to any scenarios where knowledge from different domains needs integrating. Finally, the entire system of COVIDia and its resources are publicly accessible.

ACKNOWLEDGMENT

We would like to acknowledge supports from National Natural Science Foundation of China under Grant (No. 42050105, 62202299, 62020106005, 62061146002, 61960206002), National key research and development plan No. 2022YFB3904204, Natural Science Foundation of Shanghai No. 22ZR1429100, Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University. The first author would like to thank Wu Wen Jun Honorary Doctoral Scholarship, AI Institute, Shanghai Jiao Tong University. Moreover, this work is also a contribution to the Deep-time Digital Earth (DDE) Big Science Program.

REFERENCES

[1] 2020. Docsearch. https://covid-search.doctorevidence.com/.

[2] 2023. WHO Director-General's opening remarks at the media briefing – 5 May 2023. https://www.who.int/news-room/speeches/item/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023.

[3] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. 344–354.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 722–735.

[6] Chuming Chen, Karen E. Ross, Sachin Gavali, Julie E. Cowart, and Cathy H. Wu. 2021. COVID-19 Knowledge Graph from semantic integration of biomedical literature and databases. *Bioinformatics* 37 (2021), 4597 – 4598.

[7] Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2021. LitCovid: an open database of COVID-19 literature. *Nucleic acids research* 49, D1 (2021), D1534–D1540.

[8] Qingyu Chen, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu. 2021. Artificial Intelligence (AI) in Action: Addressing the COVID-19 Pandemic with Natural Language Processing (NLP). *Annual review of biomedical data science* 4 (2021), 313–339.

[9] Meisam Dastani. 2021. An Overview of Covid-19 Dedicated Scientific Databases. *Journal of Health Literacy* 5, 4 (2021), 56–62.

[10] Joaquim de Moura, Jorge Novo, and Marcos Ortega. 2022. Fully automatic deep convolutional approaches for the analysis of COVID-19 using chest X-ray images. *Applied Soft Computing* 115 (2022), 108190.

[11] Cheng Deng, Yuting Jia, Hui Xu, Chong Zhang, Jingyao Tang, Luoyi Fu, Weinan Zhang, Haisong Zhang, Xinbing Wang, and Chenghu Zhou. 2021. GAKG: A Multimodal Geoscience Academic Knowledge Graph. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4445–4454.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Ricardo Jorge Dinis-Oliveira. 2020. COVID-19 research: pandemic versus "paperdemic", integrity, values and risks of the "speed science". *Forensic Sciences Research* 5 (2020), 174 – 187.

[14] Daniel Domingo-Fernández, Shounak Baksi, Bruce Schultz, Yojana Gadiya, Reagon Karki, Tamara Raschka, Christian Ebeling, Martin Hofmann-Apitius, and Alpha Tom Kodamullil. 2021. COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics* 37, 9 (2021), 1332–1334.

[15] Ofer Egozi, Shaul Markovitch, and Evgeniy Gabrilovich. 2011. Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)* 29, 2 (2011), 1–34.

[16] M Garcıa-Vela, J Zambrano, D Falquez, W Pincay-Musso, K Duque, N Zumba, M Barcia, J Méndez, P Valverde, P Romero-Crespo, et al. 2020. Management of virtual laboratory experiments in the geosciences field in the time of COVID-19 pandemic. In *Proceedings of ICERI2020 Conference*, Vol. 9. 8702–8711.

[17] Bernal Jimenez Gutierrez, Juncheng Zeng, Dongdong Zhang, Ping Zhang, and Yu Su. 2020. Document classification for covid-19 literature. *arXiv preprint arXiv:2006.13816* (2020).

[18] Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. In *Proceedings of EMNLP-IJCNLP: System Demonstrations*. 169–174.

[19] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020).

[20] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet* 395, 10223 (2020), 497–506.

[21] Xin Huang, Boli Chen, Lin Xiao, Jian Yu, and Liping Jing. 2021. Label-aware document representation via hybrid attention for extreme multi-label text classification. *Neural Processing Letters* (2021), 1–17.

[22] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[23] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[24] Karl E. Kim. 2021. Impacts of COVID-19 on transportation: Summary and synthesis of interdisciplinary research. *Transportation Research Interdisciplinary Perspectives* 9 (2021), 100305 – 100305.

[25] Franck Michel, Fabien L. Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena, Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, and Tobias Mayer. 2020. Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research. In *International Workshop on the Semantic Web.*

[26] Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 838–844.

[27] Simon J Porter and Daniel W Hook. 2020. How COVID-19 is changing research culture. *London: Digital Science* (2020).

[28] C Quoc and Viet Le. 2007. Learning to rank with nonsmooth cost functions. *Proceedings of the Advances in Neural Information Processing Systems* 19 (2007), 193–200.

[29] Ismael Rafols and Martin Meyer. 2010. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics* 82, 2 (2010), 263–287.

[30] Justin T Reese, Deepak Unni, Tiffany J Callahan, Luca Cappelletti, Vida Ravanmehr, Seth Carbon, Kent A Shefchek, Benjamin M Good, James P Balhoff, Tommaso Fontana, et al. 2021. KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns* 2, 1 (2021), 100155.

[31] Gerard Salton and Chris Buckley. 1988. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* 24 (1988), 513–523.

[32] Gaoussou Sanou, Véronique Giudicelli, Nika Abdollahi, Sophia Kossida, Konstantin Todorov, and Patrice Duroux. 2022. IMGT-KG: A Knowledge Graph for Immunogenetics. In *International Workshop on the Semantic Web.*

[33] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837.

[34] Bram Steenwinckel, Gilles Vandewiele, Ilja Rausch, Pieter Heyvaert, Ruben Taelman, Pieter Colpaert, Pieter Simoens, Anastasia Dimou, Filip De Turck, and Femke Ongenae. 2020. Facilitating the Analysis of COVID-19 Literature Through a Knowledge Graph. In *International Workshop on the Semantic Web.*

[35] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics*. Springer, 194–206.

[36] Zhaowei Tan, Changfeng Liu, Yuning Mao, Yunqi Guo, Jiaming Shen, and Xinbing Wang. 2016. AceMap: A novel approach towards displaying relationship among academic literatures. In *Proceedings of the 25th international conference companion on world wide web*. 437–442.

[37] Kazuya Tanaka, Riku Arakawa, Yasuaki Kameoka, and Ichiro Sakata. 2018. Recategorizing Interdisciplinary Articles using Natural Language Processing and Machine/Deep Learning. *2018 Portland International Conference on Management of Engineering and Technology (PICMET)* (2018), 1–6.

[38] Amalie Trewartha, John Dagdelen, Haoyan Huo, Kevin Cruse, Zheren Wang, Tanjin He, Akshay Subramanian, Yuxing Fei, Benjamin Justus, Kristin Persson, et al. 2020. COVIDScholar: An automated COVID-19 research aggregation and analysis platform. *arXiv preprint arXiv:2012.03891* (2020).

[39] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints* (2018), arXiv–1807.

[40] Caroline S. Wagner, J. David Roessner, Kamau Bobb, Julie Thompson Klein, Kevin W. Boyack, Joann Keyton, Ismael Rafols, and Katy Börner. 2011. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *J. Informetrics* 5 (2011), 14–26.

[41] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William Cooper Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas A. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. *ArXiv* (2020).

[42] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, et al. 2020. COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv:2007.00576* (2020).

[43] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling* 59, 9 (2019), 3692–3702.

[44] Colby Wise, Vassilis N Ioannidis, Miguel Romero Calvo, Xiang Song, George Price, Ninad Kulkarni, Ryan Brand, Parminder Bhatia, and George Karypis. 2020. COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv:2007.12731* (2020).

[45] Yang Yang, Na Zhao, Ting Ma, Ze Yuan, and Cheng Deng. 2022. 'Paperdemic' during the COVID-19 pandemic. *European Journal of Internal Medicine* (2022).

[46] Dejiao Zhang, Feng Nan, Xiaokai Wei, Shangwen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. *arXiv preprint arXiv:2103.12953* (2021).