2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS) | 978-0-7381-1180-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/SNAMS52053.2020.9336541

Cone-KG: A Semantic Knowledge Graph with News Content and Social Context for Studying Covid-19 News Articles on Social Media

Feras Al-Obeidat*, Oluwasegun Adedugbe†, Anoud Bani Hani*, Elhadj Benkhelifa†, Munir Majdalawieh*

*College of Technological Innovation Zayed University
United Arab Emirates

Email: Feras.Al-Obeidat,anoud.Bani-Hani,Munir.Majdalawieh@zu.ac.ae

†Cloud Computing Applications & Research Lab, School of Computing Staffordshire University
Stoke-on-Trent, United Kingdom

Email: e.benkhelifa@staffs.ac.uk, oluwasegun.adedugbe@research.staffs.ac.uk

***Abstract*—Semantic knowledge graphs provide very significant benefits for structuring and analysing huge amounts of aggregated data across diverse heterogeneous sources. Beyond quick and efficient data query and analysis, they facilitate inference from data and generation of insights for several purposes. With the multi-faceted global challenges posed by the COVID-19 pandemic, this research focused on the use of a semantic knowledge graph to model, structure and store COVID-related news articles centrally and semantically towards knowledge discovery, knowledge acquisition and advanced data analytics for understanding varying metrics relating to the virus towards curbing its spread. The semantic knowledge graph provides a platform for researchers, data analysts and data scientists across societal sectors to investigate and recommend strategies towards addressing the challenges it poses to the global society.**

Keywords—Knowledge Graphs, Semantic Graphs, Semantic Web, COVID-19 News, Social Media, Social Data Analysis.

I. INTRODUCTION

Semantic knowledge graphs provide a means of knowledge representation by generating nodes and edges from a corpus of underlying data usually aggregated from diverse heterogenous sources. This facilitates interesting relationships between entities such as words, phrases and derived concepts, enabling extraction of vital insights and inferences based on the data schema. They provide a description of real-world entities with interrelations, arranged in a graphical structure [1]. In addition, they cover a variety of topical domains, providing definitions for classes and relations of entities, thereby fostering interrelations between them. Furthermore, they provide context awareness and are created based on one of two major approaches; top-down or bottom-up approaches. The top-down approach defines an ontology and its schema, followed by addition of knowledge instances into the knowledge base. Its emphasis is on well-defined domain ontologies for representation of actual instances of knowledge graphs. The bottom-up approach on the other hand focuses on extraction of knowledge instances from the Linked Open Data (LOD) or other knowledge resources. Utilising the top-level approach through knowledge instances leverages knowledge fusion of the populated instances [2]. The bottom-up approach's knowledge construction process is iterative with updates. The process includes knowledge acquisition, fusion, storage and retrieval. Knowledge is acquired from major sources which include structured, semi-structured and unstructured data. In the process of extracting knowledge, the three main features extracted are entities, attributes and relations. The knowledge fusion process takes place iteratively and requires constant construction of ontology and evaluation of its quality. Cone-KG is a COVID-19 knowledge graph which contains over 6 million COVID-19 related news articles from diverse publishers and authors with over 100 million statements in the ensuing semantic graph database, providing an infrastructure for social data analytics relating to the covid-19 pandemic which can be very beneficial for researchers, data scientists and organisations. The knowledge graph also facilitates other significant benefits and application areas based on predictive analytics, such as better engagement with the public and their sentiments or opinions, improved private/public relations and access to vital data. It also enhances leveraging artificial intelligence solutions such as personal assistants or chatbots as well as monitoring multiple metrics for improved public health. The planning, management and optimisation of multiple data resources is another benefit from the KG (Knowledge Graph). The remaining sections of the paper are as follows: section 2 provides some background to the domain, section 3 reviews several related works, section 4 describes implementation methodology for the knowledge graph while section 5 analyses it, with some use case scenarios. Section 6 concludes the paper, with recommendations for further work in the area.

II. BACKGROUND

Efficient querying and merge-join of triple patterns is one of the merits of RDF-based knowledge graph storage. The storage of data as structured nodes, edges and properties of graphs enables effective querying and traversal of the data as well as supporting various graph mining algorithms. A renowned example of graph database is Neo4j which is open source and offers native graph storage. Alongside SPARQL for querying large-scale knowledge graphs via an endpoint, other languages include Cypher [3], Gremlin [4], and G-CORE [5]. The output formats of SPARQL query results include JSON, JSON-LD, XML, RDF/XML, RDF/N3, CSV, TSV and HTML, with the majority being machine-readable. However, certain formats of query results of knowledge

graphs are text-based. Data visualisation is also enabled via browsers such as IsaViz, RDF Gravity, DBpedia Mobile, Fenfire and OpenLink Data Explorer [2]. In comparison to other forms of structured data storage, knowledge graphs provide diverse layers of data to meet a wide range of data processing and analytical needs. These layers can be categorised as a data layer which presents aggregated heterogenous data from diverse sources. An information layer provides schema definition and data structuring based on the schema. Knowledge layer presents data as things, represented within a semantic (ontology-driven) graph database, such as Neo4j, GraphDB, Virtuoso and AWS Neptune while a wisdom layer provides inferencing and reasoning derived from represented knowledge within the knowledge graph [6]. There are several benefits of KG implementations. They have the capability for combining different data types, extracting relationships, and discovering patterns which further enable them to infer new facts from existing datasets. For example, it is possible to traverse, explore, visualize and infer important insights from a KG built using linked data of different domains such as, social, commerce, life science, geographic. KGs graph-based structure also makes them very flexible structurally. Other forms of representation can be easily mapped to graphs. In graphs, links can be effectively traversed to determine how different parts of a domain relate to each other. Moreover, within KGs, the meaning of data is embedded in the form of ontology along with the data in the graph, supporting natural language-based query for data extraction [7]. Furthermore, graph-related algorithms and computing techniques such as shortest path discovery can be applied to KGs, which further increase its intelligence. KGs are also easily extendable, being scalable for additional, automated data storage and linkage. This makes them well suited for dynamic, data-intensive applications with continuous stream of data across disparate sources. They also support provenance, using reinforcement statements which preserve metadata of data, in this case such as article publisher, author and date; vital data for credibility verification.

III. Related Work

Since the outbreak of the COVID-19 pandemic, there has been immense data about the subject with diverse datasets focusing on specific types of elements. While some others focus on similar elements, the perspective and actual data differ. With the disparate nature of these data in terms of format, schema, location and several other metrics, it becomes challenging to aggregate them and establish links and connections between them. A holistic approach such as utilising knowledge graphs for aggregating, structuring and storage of such data is required for efficient use. The deployment of semantic knowledge graphs to make meaning of and contextualise these data in RDF formats can be observed across various research efforts. This is prominent in healthcare for assisting researchers and practitioners when seeking to evaluate established facts and knowledge about the disease. Health organizations and governments have embraced use of knowledge graphs to hasten research and find a basis for decisions relating to treatment and preventive measures. Furthermore, there appears a great need to progressively mesh new COVID-19 and SARS-CoV-2 virus information with the rich collective data that already exists in the massive LOD Cloud Knowledge Graph. In addition, cost and accuracy are critical benefits drawn from the use of knowledge graphs for COVID-19 research. The work by [8] presented a retriever-ranker semantic search engine known as CO-Search towards handling complex queries over literature related to COVID-19. It was designed to ease the burden of health workers in seeking scientific answers when required. The retriever was based on a SiameseBERT encoder [9] characterised by linear composition of a TF-IDF vectorizer [10] and reciprocal-rank fusion [11] with a BM25 vectorizer. It also included a multi-hop question-answering module in the ranker for adjustment of retriever scores in collaboration with a multi-paragraph abstractive summarizer [12]. A bipartite graph of document paragraphs and citations was generated, with 1.3 million tuples to train the encoder towards accounting for the domain-specific and relatively limited dataset. The proposed CO-Search was evaluated using data from TREC-COVID information retrieval challenge [13]. The use of significant metrics such as normalised discounted cumulative gain, precision, mean average precision and binary preference for effective performance on such datasets is imperative.

Similarly, [14] developed a new and all-inclusive knowledge discovery framework known as COVID-KG towards the extraction of fine-grained multimedia knowledge elements, including entities, relations and events from scientific literature. The system reads existing papers to construct a multimedia knowledge graph. The information extraction (IE) is made up of coarsely-grained entity extraction and entity linking, utilising entity ontology defined in the Comparative Toxicogenomics Database (CTD), thereby obtaining a Medical Subject Headings (MeSH) Unique ID for each mention. All defined entities were linked to the CTD and extracted subtypes of relations. Through event extraction, they extracted event types and entities' roles. The researchers applied a fine-grained entity extraction system, called CORD-NER for the extraction of entities towards enriching the KG. Furthermore, a visual IE subsystem developed focused on extracting visual information from figure images and enriching the knowledge graph. Figure-separator helped in detecting and separating all non-overlapping image regions. The study of [15] presented a COVID-19 knowledge graph, CKG with a wide-range towards extracting and visualizing complex relationships between COVID-19 scientific articles. The researchers used the data latent schema to construct CKG, followed by enrichment with biomedical entity information extracted from the unstructured text of articles using scalable AWS technologies to form relations in the graph. After the construction of CKG using CORD-19 datasets, a document similarity engine was developed combining both semantic and relationship information from CKG. The training of a topic model on the corpus and Amazon Comprehend Medical service extracted biomedical entity relationships and highly abstracted topics from the unstructured text of articles. SciBERT generated semantic embeddings for each article while Knowledge Graph Embedding (KGE) and Graph Neural Network technologies generated embeddings for entities and relations of CKG.

The analysis of public Covid-19 datasets such as the Harvard INDRA COVID-19 Knowledge Network (INDRA CKN) dataset, the Blender lab COVID knowledge graph dataset (Blender KG), and COVID-19 Open Research Dataset (CORD-19) can also be observed to have been based on semantic visualisation strategies [16]. With Integrated Network and Dynamical Reasoning Assembler (INDRA)

utilised in assembling extracted events, and NER models trained on the BIONLP 13KG corpus, [17] utilised Elasticsearch for back-end data index and data visualisation via Kibana. In addition, the CORD19STS dataset was developed via annotation of sentence pairs retrieved from CORD-19 challenge using various sampling techniques to generate one million sentence pairs and a fine-tuned BERT-like language model called Sen-SCI-CORD19-BERT towards calculation of similarity scores between sentence pairs in providing a balanced dataset based on semantic similarity levels [17]. The research further employed SCI-BERT to model textual information and fine-tuned SCI-BERT over pre-processed CORD-19 text through Masked Language Modelling to produce the language model known as SCI-CORD19-BERT. The combination of SCI-CORD19-BERT with a Siamese network architecture created Sen-SCI CORD19-BERT. [18] formalised and extracted insights from the dataset introduced in the study of [19] and the CORD-19 dataset towards identification of experts and bio-entities related to COVID-19, employing various machine learning, deep learning, and knowledge graph construction and mining strategies. The BioBERT model was utilised for the recognition of entities while fine-tuning was on the NCBI disease dataset. The co-occurrence frequency-based knowledge graph was built using Gelphi. As a cosine similarity knowledge-based graph, the entities were firstly normalized based on a ruleset to deal with entity name disambiguation challenges while Word2Vec facilitated conversion of the normalized entities to vector with length of 100.

From the diverse related work, it can be observed that while semantic technologies in the form of knowledge graphs have been applied towards COVID19-related topics, its use for COVID-19 news articles from both online news publishers and social media has not been exploited at large scale, hence defining a novel niche for this research. The Cone-KG is beneficial in diverse ways such as knowledge discovery, knowledge acquisition and articles-based analytics towards better understanding of various features related to COVID-19 and how its spread can be curbed. Some of such features include its diffusion, propagation and sustenance models.

IV. Implementation

The data we used for experimental purposes in this paper were COVID-19 news articles curated by third parties from diverse news publishing portals on the web and mostly shared on social media. The total size of data collected in different formats was about 35GB. Table 1 presents the different data sources, data format, number of articles in each dataset and their sizes.

TABLE I. Raw Datasets used for Cone-KG Construction

Dataset	Format	# Articles	Size
Aylien Covid [20]	JSON	1,200,000	17.84 GB
IEEE Dataport [21]	JSON	5,200,000	13.70 GB
Covid-19 Public Media Dataset [22]	CSV	200,000	3.7 GB

The COVID-19 Public Media Dataset articles bundle consists of four datasets. These contain data about COVID-19 related articles; date of publication, publisher, author, title, topic area (such as general, finance and science) and the news content. Each cell was covered in the Cone-KG model. That is, all columns were mapped to the model. The Aylien COVID-19 dataset was obtained from Aylien NLP [20] with over 1.2 million curated COVID-19 news articles. While this dataset largely contained same information as the previously mentioned datasets, it had some additional parameters such as image, image URL and type. The IEEE Dataport[22] dataset also significantly overlapped with the previous two datasets. However, it had several other features which further enhanced the social context representation for each article. These included country, social media shares count (on Facebook, Twitter, etc.), sentiments, persons, external links, persons, comments and many more.

A. Data Extraction

To extract relevant data from each of the datasets, we implemented efficient and scalable parsers for the data types i.e., JSON, JSONL, CSV, TXT. For the JSON files specifically and due to their large sizes, we implemented JSON streaming parsers on Amazon AWS cloud to extract the relevant data after careful study of the data schema. The JSON streaming parser enables the JSON strings to be read into memory in bits rather than all at once. This required implementing a JSON streaming parser "Listener" interface which is then passed into the parser, enabling the "Listener" to receive events from the streaming parser. Based on these, we could access both object properties and array elements within the JSON files and retrieve them iteratively. The extracted data were stored in MySQL database tables which had schemas matching those of the JSON files. With relevant data extracted and structured in the tables, they were exported in CSV format for transformation to RDF data.

B. Data Cleaning and Processing

Before we RDFized the CSV data, the data was observed to be very noisy and some objects or entities which should be represented as RDF resources were having special characters and spaces which required removal. Thus, we solved those issues so that the generated resources in RDF can be valid URIs. In addition, as the data were collected from different sources, we found date formats were different, therefore, we changed those into a common format understandable by RDF query language SPARQL. A new column was also added in each dataset as "author-new" (copied from the author column). In the new column each space and special characters were replaced by dash. This was done because each author name is used as a resource in RDF data and RDF resources should be valid URIs. Many author cells were empty in each dataset. These were replaced with "unknown-id" where "id" is the number corresponding to row number. This can tell us how many articles were published by unknown authors. Existing URLs for images and domains in each dataset were also validated and changed where required.

C. Data Transformation and KG Construction

After the data were cleaned, they were transformed to RDF using our RDFizer tool. To RDFize COVID-19 data, we modelled it using Fandet ontology (available via https://github.com/rif-zu/fandet-ontology). For ease of importing and storage, the actual data were broken into smaller chunks before transformation to RDF format. This also facilitated ease of fitting into memory. A configuration Python script was utilised to define input and output

directories as well as prefix of the URIs for the triples. Thereafter, a structure was defined for how the RDF entities look like and how the URIs are built. Each file is then read from the input directory, parsed and the RDF data generated. An excerpt of data modelled using the ontology is presented in Figure 1.

```
@prefix ns0: <http://www.semanticweb.org/developer/ontologies/2020/3/Fandet_Ontology#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://www.fandet.com/data/event/777>
 a
<http://www.semanticweb.org/developer/ontologies/2020/3/Fake_news_Ontology#Covid19_News_Transmission_Event> ;
 ns0:hasPublisher <http://www.fandet.com/data/publisher/Alan-Brochstein> ;
 ns0:hasContent <http://www.fandet.com/data/social-media-object/777> .

<http://www.fandet.com/data/publisher/Alan-Brochstein>
 a ns0:Social_Media_Agent ;
 ns0:hasName "Alan Brochstein" .

<http://www.fandet.com/data/social-media-object/777>
 a ns0:Social_Media_Object ;
 ns0:hasPublishedDate <http://www.fandet.com/data/date/777> .

<http://www.fandet.com/data/date/777>
 a ns0:Date ;
 ns0:hasDate "05/04/2020"^^xsd:dateTime .

<http://www.fandet.com/data/title/777>
 a ns0:Title ;
 ns0:hasDescription "Take These Steps To Help Prevent The Coronavirus From Infecting Your Stocks" .
```

Fig. 1. An example of modelled data within the Cone knowledge graph.

The knowledge graph was created using mappings based on a sub-section of Fandet ontology as shown in Figure 2. From the figure, it can be observed that "Covid19-News-Transmission-Event" represents any published event online with social media sharing capabilities with the publisher mapped in the model as "Social_Media_Agent". Every social media agent has its name. Every published event has its contents such as data and metadata; represented by a class "Social_Media_Object" and its associated properties.

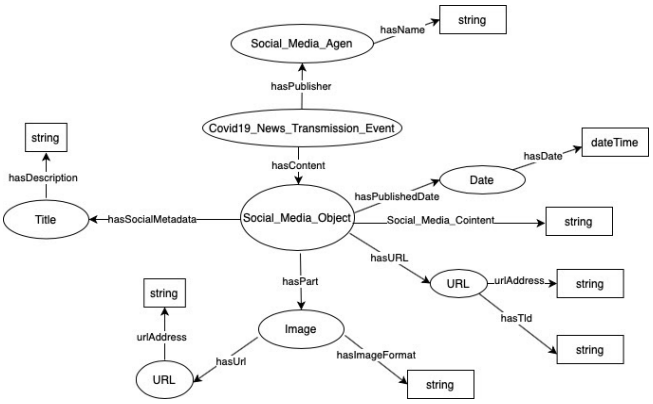

Fig. 2. Cone Knowledge Graph Model

As an example, Table II shows a SPARQL query on the database that verifies mapping according to the model given in Figure 1. The query shows what data a particular publisher has in our created KG. News_Transmission_Events are the fundamental concepts that are modelled in this ontology, where news is any kind of content that is being transmitted on social media. However, they are not only characterised by some content but also by the context of occurrence. Social_Media_Agent is a high-level concept that denotes different types of agents on social media. In the context of a News_Transmission_Event, a social media agent encounters or promotes news content. Social_Media_Object is also a high-level concept representing information that is published on a social media platform. The concept is further specified by its subclasses.

TABLE II. SPARQL QUERY FOR A SPECIFIC AUTHOR FROM THE KNOWLEDGE GRAPH

```
PREFIX fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fandet_
_Ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
select * {

?covidEvent a ?EventType.
?covidEvent fno:hasPublisher ?publisher.
?publisher fno:hasName ?name.

?covidEvent fno:hasContent ?sMediaObject.
?sMediaObject fno:hasUrl ?Url.
?Url fno:urlAddress ?urlAddress.
?Url fno:hasTld ?TldName.
?sMediaObject fno:socialMediaContent ?sMediaContent.
?sMediaObject fno:partOf ?topic.

?sMediaObject fno:hasSocialMetaData ?titleSMetaData.
?titleSMetaData fno:hasDescription ?titleDesc.

?sMediaObject fno:hasPublishedDate ?sMediaPublishedDate.
?sMediaPublishedDate fno:hasDate ?publishedDate.

FILTER (?publishedDate ="2020-01-26"^^xsd:dateTime)
VALUES ?publisher { <http://www.fandet.com/data/publisher/Sean-Martin> }
}
```

D. *Semantic Graph Database*

To load RDF data into triplestore, GraphDB was utilised, running as a standalone server with a preconfigured web server. The size of RDF data as total distinct triples, subjects, predicates, and objects are shown in Table III.

TABLE III. RDF DATA SIZE IN COUNTS FOR TRIPLES, SUBJECTS, PREDICATES AND OBJECTS

Triples	Subjects	Predicates	Objects
101,467,747	35,016,070	21	45,272,480

With GraphDB, repositories are created for each RDF data and then the actual data files can be imported into the repository. The data import can be based on "User Data", referring to data files from a user-defined directory remote to the server or based on "Server Files", in which case the files are fetched from specific directories on the server. In this case, "Server Files" were utilised considering their significantly large sizes. Generally, RDF data loading can also be via SPARQL or RDF4J APIs as well as via the GraphDB LoadRDF tool. Once successfully loaded, the data can be explored in diverse ways such as exploring instances, similarity indexes, visualising class hierarchies and class relationships. Visual graphs can also be configured based on user-defined SPARQL queries. Table IV presents dependencies of our KGs main classes and relationships between them, in terms of count for both incoming and outgoing links. The knowledge graph is publicly available and can be accessed via: https://github.com/rif-zu/cone-knowledge-graph

TABLE IV. CONE-KG CLASS DEPENDENCIES SHOWING RELATIONSHIPS BETWEEN MAIN CLASSES

Class Names	Number of Links
Fno:Social_Media_Object	25.88 million
Fno:Url	10.75 million
Fno:COVID19_News_Transmission_Event	8.28 million
Fno:Image	7.79 million
Fno:Date	4.14 million
Fno:Title	4.14 million
Fno:Social_Media_Agent	4.14 million
Fno:Social_Metadata	2.71 million

V. KNOWLEDGE GRAPH ANALYSIS

As our created KG comprises data about COVID-19 which was collected from different sources of information. We explored our KG to find hidden insights which would be hard to find or even not possible with relational databases. We performed different types of analytics by executing various SPARQL queries. For example, we can find which particular author has published articles about COVID-19 through multiple news publishers. Also, based on free text search using SPARQL queries we can find how many articles within the different datasets includes specific keywords such as COVID, Coronavirus, Hydroxychloroquine or SARS-CoV-2. When we downloaded data from the different sources, it was observed that some datasets have more parameters than others. For example, country parameter was available in one and missing in another. After creating KG we were able to infer the country in other datasets based on the common publishers. Some of the other analytics we performed are described in the following sub-sections.

A. Find all distinct COVID-19 news articles publishers

This resulted into a result set of 1,076,612 publishers across the different datasets, which suggested a very high number of publishers have been involved globally. However, as some of the articles from the raw datasets didn't have a publisher's name defined in which case, we labelled such as "unknown" followed by a hyphen and an integer number, it implies a potential level of repetition for publishers. In addition, it was observed from the raw datasets that publishers and authors were not very distinctly defined, hence, the result set here had both publishers and authors listed. The SPARQL query for this is presented in Table V.

TABLE V. SPARQL QUERY FOR DISTINCT COVID-19 NEWS ARTICLES PUBLISHERS FROM THE KG

```
PREFIX fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fandet_Ontology#>

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

select distinct ?name where {

?publisher fno:hasName ?name. }
```

B. Articles count by major COVID-19 news publishers

The goal here was to know number of COVID-19 related articles by top news publishing sites. The SPARQL query is based on a keyword search; retrieving any article with COVID, Coronavirus or any other related keywords in its content. From the results obtained, they were over 1,000 different publishers. Hence, we defined a limit for the top 10 in articles count. The results revealed "reuters.com" as top publisher with 69,609 of such articles in our knowledge graph and as at the time of writing, it was observed that the portal had over 74,000 articles related to the subject matter, which revealed that they had published some more articles on the topic after curation of the primary datasets utilised in this research. Table VI presents the top 10 publishers as revealed by the KG. It is noteworthy to mention though, that the results are only based on the data we have which is not representative of all available data on the subject matter, hence, there can be variations in actual values if all available data were curated and stored in a single knowledge graph.

TABLE VI. SPARQL QUERY AND RESULT SET FOR TOP 10 COVID-19 NEWS PUBLISHERS

```
PREFIX fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fandet_Ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    select (count(distinct ?urlAddress) as ?articles) ?channel where {
      ?covidEvent a ?EventType;
      fno:hasContent ?sMediaObject.
      ?sMediaObject fno:hasUrl ?Url.
      ?Url fno:urlAddress ?urlAddress.
      ?Url fno:hasTld ?TldName. BIND (?TldName as ?channel).
} group by ?channel order by DESC (?articles)
```

Channel/News Publisher	# Articles
"reuters.com"	69,609
"dailymail.co.uk"	67,989
"yahoo.com"	57,811
"business-standard.com"	34,728
"indiatimes.com"	31,568
"urdupoint.com"	30,082
"bizjournals.com"	25,936
"cbslocal.com"	21,945
"express.co.uk"	21,468
"onenewspage.com"	20,960

C. Count of authors for specific news publishers

In this case, the result set had "yahoo.com" with highest count on number of authors at 23,976. However, we discovered as well that we had designed our knowledge graph to treat domains and subdomains as different entities. Hence, "yahoo.com" is treated as a distinct entity from its subdomain "news.yahoo.com". Our attention was really drawn to this as the result set had each of these two in the top 10 results. We also discovered that other "yahoo.com" subdomains likewise had very high counts, such as "finance.yahoo.com" which had 6,454 authors, even though it was not within the top 10. A similar scenario was observed for some others, such as "news.com". Furthermore, entities such as "Google" with diverse top-level domains such as ".com", ".ca" and ".co.uk" were all treated as independent entities. Table VII presents the top 10 from this result set and the corresponding SPARQL query while some other analytics based on different SPARQL queries executed are presented in Table VIII.

TABLE VII. SPARQL QUERY AND RESULT SET FOR TOP AUTHOR COUNT PER COVID-19 NEWS PUBLISHERS

```
PREFIX fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fandet_Ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    select (count(distinct ?publisher) as ?publishers) ?channel where {
```

```
        ?covidEvent a ?EventType.
        ?covidEvent fno:hasPublisher ?publisher.
        ?covidEvent fno:hasContent ?sMediaObject.
        ?sMediaObject fno:hasUrl ?Url.
        ?Url fno:urlAddress ?urlAddress.
        ?Url fno:hasTld ?TldName. BIND (?TldName as ?channel).
} group by ?channel order by DESC (?publishers)
```

Channel/News Publisher	# of Authors
"yahoo.com"	23,976
"indiatimes.com"	20,987
"reuters.com"	16,332
"news18.com"	16,227
"news.yahoo.com"	12,725
"onenewspage.com"	12,446
"cnbc.com"	11,999
"globalbankingandfinance.com"	10,924
"sharenet.co.za"	10,148
"google.ca"	9,090

TABLE VIII. SOME OTHER SPARQL QUERIES EXECUTED ON THE KNOWLEDGE GRAPH

#1 Total count of news publishers within the KG = 52260

```
PREFIX fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fake_news_Ontology#>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

    select (count(distinct ?TldName)  as ?channels)  where {

        ?covidEvent a ?EventType.
        ?covidEvent fno:hasPublisher ?publisher.

        ?covidEvent fno:hasContent ?sMediaObject.
        ?sMediaObject fno:hasUrl ?Url.
        ?Url fno:urlAddress ?urlAddress.
        ?Url fno:hasTld ?TldName.
}
```

#2 How many distinct articles by a specific author (Sean Martin) = 161

```
prefix fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fake_news_Ontology#>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

    select (count(distinct ?urlAddress) as ?distinctArticlesAddresses)
where {
        ?covidEvent a ?EventType.
        ?covidEvent fno:hasPublisher ?publisher.

        ?covidEvent fno:hasContent ?sMediaObject.
        ?sMediaObject fno:hasUrl ?Url.
        ?Url fno:urlAddress ?urlAddress.
    VALUES ?publisher { <http://www.fakenews.com/data/publisher/Sean-Martin>}
}
```

#3 Name of publisher(s) for a specific article

```
prefix fno:
<http://www.semanticweb.org/developer/ontologies/2020/3/Fake_news_Ontology#>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

    SELECT ?g ?publisher ?name ?TldName
    {

      ?covidEvent a ?EventType.
      ?covidEvent fno:hasPublisher ?publisher.
      ?publisher fno:hasName ?name.

      ?covidEvent fno:hasContent ?sMediaObject.
      ?sMediaObject fno:hasUrl ?Url.
      ?Url fno:urlAddress
"https://www.forbes.com/sites/judystone/2019/12/12/how-prepared-are-we-for-the-next-pandemic-not-very-experts-show/".
      ?Url fno:hasTld ?TldName. }
```

VI. CONCLUSIONS AND FUTURE WORK

With the multi-faceted global challenges posed by the COVID-19 pandemic, it is observed that a plethora of information related to it has been published across diverse news portals and social media. However, these data exists in different formats such as text, CSV, JSON and XML. Our main aim was to provide a common model for this disrupted information that can provide a common platform for researchers, data analysts and data scientists across societal sectors to investigate, discover and recommend strategies towards addressing the challenges of COVID-19. With knowledge graphs having garnered significant attention from both industry and academia and becoming indispensable in scenarios where data is diverse and in large-scale, we modelled the collected COVID-19 data as "directed label graphs" called Linked Data using Fandet ontology. One of the motivations behind creating the semantic knowledge graph was because of the schema-less representation and data. It has been observed in relational databases that with the increase of data particularly where they are gathered from different sources, it becomes difficult or costly to handle and remodel the underlying schema for relational databases. As an attempt, in this paper, we presented a KG consisting of information related to COVID-19. We also presented an analytical overview of the generated KG and tried to dig deeper. We executed different SPARQL queries to find insights from our KG. The KG is made available via an online repository for further exploration by the research community towards COVID-19 related insights. Finally, we plan to make further releases of the KG by enriching from different data sources so it can always include updated news articles relating to the pandemic. We also have a plan to integrate our knowledge graph with other publicly available ones on the web.

REFERENCES

[1] Paulheim, H., 2017, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic Web Journal 8, 3 (2017), 489–508. https://doi.org/10.3233/SW-160218

[2] Zhao, Z., Han, S., and So, I., 2018, Architecture of Knowledge Graph Construction Techniques, International Journal of Pure and Applied Mathematics, Volume 118 No. 19 2018, 1869-1883.

[3] Francis, N., et al., 2018, Cypher: An Evolving Query Language for Property Graphs, See [112], 1433–1445.

[4] Rodriguez, M. A., 2015, The Gremlin graph traversal machine and language, In Proceedings of the 15th Symposium on Database Programming Languages, Pittsburgh, PA, USA, October 25-30, 2015, James Cheney and Thomas Neumann (Eds.). ACM Press, 1–10.

[5] Angles, R. et al, 2018, G-CORE: A Core for Future Graph Query Languages, See [112], 1421–1432.

[6] Yuan, J., Jin, Z., Guo, H., Jin, H., Zhang, X., Smith, T., & Luo, J. (2020). Constructing biomedical domain-specific knowledge graph with minimum supervision. Knowledge and Information Systems, 62(1), 317-336.

[7] Rizun, M. and Meister, V.G., 2017, September. Analysis of Benefits for Knowledge Workers Expected from Knowledge-Graph-Based Information Systems. In EuroSymposium on Systems Analysis and Design (pp. 25-39). Springer, Cham.

[8] Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D. and Socher, R., 2020. Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv preprint arXiv:2006.09595.

[9] Reimers, N. and Gurevych. I., 2019, Sentence-bert: Sentence embeddings using Siamese bertnetworks, arXiv preprint arXiv:1908.10084.

[10] Shahmirzadi, O., Lugowski, A., and Younge, K., 2019, Text similarity in vector space models: a comparative study, In ICMLA 2019, pages 659–666. IEEE.

[11] Cormack, G. V., Clarke, C. L. A., and Buettcher, S., 2019, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, In SIGIR 2009, pages 758–759.

[12] Asai, A., Hashimoto, K., Hajishirzi, H., Socher, R., and Xiong, C., 2020, Learning to retrieve reasoning paths over wikipedia graph for question answering, In ICLR 2020.

[13] Voorhees, E. et al., 2020, TREC-COVID: Constructing a pandemic information retrieval test collection, arXiv preprint arXiv:2005.04474, 2020.

[14] Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, H., Liu, W. and Chauhan, A., 2020. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. arXiv preprint arXiv:2007.00576.

[15] Wise, C., Ioannidis, V.N., Calvo, M.R., Song, X., Price, G., Kulkarni, N., Brand, R., Bhatia, P. and Karypis, G., 2020. COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature. arXiv preprint arXiv:2007.12731.

[16] Tu, J., Verhagen, M., Cochran, B. and Pustejovsky, J., 2020. Exploration and Discovery of the COVID-19 Literature through Semantic Visualization. arXiv preprint arXiv:2007.01800.

[17] Guo, X., Mirzaalian, H., Sabir, E., Jaiswal, A. and Abd-Almageed, W., 2020. CORD19STS: COVID-19 Semantic Textual Similarity Dataset. arXiv e-prints, pp.arXiv-2007.

[18] Chen, C., Ebeid, I.A., Bu, Y. and Ding, Y., 2020. Coronavirus Knowledge Graph: A Case Study. arXiv preprint arXiv:2007.10287.

[19] Dernoncourt, F. and Lee, J. Y., 2017, Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv preprint arXiv:1710.06071.

[20] Aylien News API. 2020. Free Coronavirus News Dataset. [online] Available at: <https://aylien.com/blog/free-coronavirus-news-dataset> [Accessed 15 May 2020].

[21] Rabindra Lamsal, March 13, 2020, "Coronavirus (COVID-19) Tweets Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/781w-ef42.

[22] Lipenkova, J., 2020. COVID-19 Public Media Dataset. [online] Kaggle. Available at: <https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset> [Accessed 15 May 2020].