Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research

Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, Tobias Mayer, et al.

Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research

Franck Michel, Fabien Gandon, Valentin Ah-Kane, Anna Bobasheva, Elena Cabrio, Olivier Corby, Raphaël Gazzotti, Alain Giboin, Santiago Marro, Tobias Mayer, Mathieu Simon, Serena Villata, and Marco Winckler

University Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
franck.michel@cnrs.fr, fabien.gandon@inria.fr, valentin.ah-kane@etu.univ-cotedazur.fr, anna.bobasheva@inria.fr, elena.cabrio@inria.fr, olivier.corby@inria.fr, raphael.gazzotti@inria.fr, alain.giboin@inria.fr, santiago.marro@inria.fr, tobias.mayer@inria.fr, mathieu.simon@inria.fr serena.villata@inria.fr, winckler@i3s.unice.fr

Abstract. Scientists are harnessing their multi-disciplinary expertise and resources to fight the COVID-19 pandemic. Aligned with this mindset, the Covid-on-the-Web project aims to allow biomedical researchers to access, query and make sense of COVID-19 related literature. To do so, it adapts, combines and extends tools to process, analyze and enrich the "COVID-19 Open Research Dataset" (CORD-19) that gathers 50,000+ full-text scientific articles related to the coronaviruses. We report on the RDF dataset and software resources produced in this project by leveraging skills in knowledge representation, text, data and argument mining, as well as data visualization and exploration. The dataset comprises two main knowledge graphs describing (1) named entities mentioned in the CORD-19 corpus and linked to DBpedia, Wikidata and other BioPortal vocabularies, and (2) arguments extracted using ACTA, a tool automating the extraction and visualization of argumentative graphs, meant to help clinicians analyze clinical trials and make decisions. On top of this dataset, we provide several visualization and exploration tools based on the Corese Semantic Web platform, MGExplorer visualization library, as well as the Jupyter Notebook technology. All along this initiative, we have been engaged in discussions with healthcare and medical research institutes to align our approach with the actual needs of the biomedical community, and we have paid particular attention to comply with the open and reproducible science goals, and the FAIR principles.

Keywords: COVID-19, arguments, visualization, named entities, linked data

1 Bringing COVID-19 data to the LOD: deep and fast

In March 2020, as the Coronavirus infection disease (COVID-19) forced us to confine ourselves at home, the Wimmics research team[1] decided to join the effort

[1] https://team.inria.fr/wimmics/

of many scientists around the world who harness their expertise and resources to fight the pandemic and mitigate its disastrous effects. We started a new project called *Covid-on-the-Web* aiming to make it easier for biomedical researchers to access, query and make sense of the COVID-19 related literature. To this end, we started to adapt, re-purpose, combine and apply tools to publish, as thoroughly and quickly as possible, a maximum of rich and actionable linked data about the coronaviruses.

In just a few weeks, we deployed several tools to analyze the *COVID-19 Open Research Dataset* (CORD-19) [20] that gathers 50,000+ full-text scientific articles related to the coronavirus family. On the one hand, we adapted the ACTA platform[2] designed to ease the work of clinicians in the analysis of clinical trials by automatically extracting arguments and producing graph visualizations to support decision making [13]. On the other hand, our expertise in the management of data extracted from knowledge graphs , both generic or specialized, and their integration in the HealthPredict project [9,10], allowed us to enrich the CORD-19 corpus with different sources. We used DBpedia Spotlight [6], Entity-fishing[3] and NCBO BioPortal Annotator [12] to extract Named Entities (NE) from the CORD-19 articles, and disambiguate them against LOD resources from DBpedia, Wikidata and BioPortal ontologies. Using the Morph-xR2RML[4] platform, we turned the result into the *Covid-on-the-Web RDF dataset*, and we deployed a public SPARQL endpoint to serve it. Meanwhile, we integrated the Corese[5] [5] and MGExplorer [4] platforms to support the manipulation of knowledge graphs and their visualization and exploration on the Web.

By integrating these diverse tools, the Covid-on-the-Web project (sketched in Fig. 1) has designed and set up an integration pipeline facilitating the extraction and visualization of information from the CORD-19 corpus through the production and publication of a continuously enriched linked data knowledge graph. We believe that our approach, integrating argumentation structures and named entities, is particularly relevant in today's context. Indeed, as new COVID-19 related research is published every day, results are being actively debated, and moreover, numerous controversies arise (about the origin of the disease, its diagnosis, its treatment...) [2]. What researchers need is tools to help them get convinced that some hypotheses, treatments or explanations are indeed relevant, effective, etc. Exploiting argumentative structures while reasoning on named entities can help address these user's needs and so reduce the number of controversies.

The rest of this paper is organized as follows. In Section 2, we explain the extraction pipeline set up to process the CORD-19 corpus and generate the RDF dataset. Then, Section 3 details the characteristics of the dataset and services made available to exploit it. Sections 4 and 5 illustrate the current exploitation and visualization tools, and discuss future applications and potential impact of the dataset. Section 6 draw a review of and comparison with related works.

[2] http://ns.inria.fr/acta/

[3] https://github.com/kermitt2/entity-fishing

[4] https://github.com/frmichel/morph-xr2rml/

[5] https://project.inria.fr/corese/

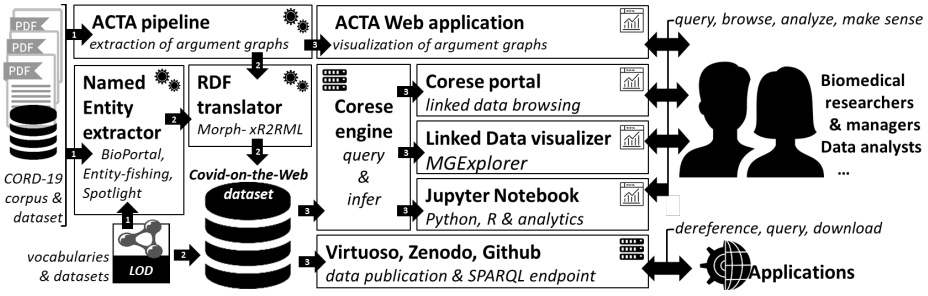

Fig. 1. Covid-on-the-Web overview: pipeline, resources, services and applications

2 From CORD-19 to the Covid-on-the-Web Dataset

The COVID-19 Open Research Dataset [20] (CORD-19) is a corpus gathering scholarly articles (ranging from published scientific publications to pre-prints) related to the SARS-Cov-2 and previous works on the coronavirus family. CORD-19's authors processed each of the 50,000+ full text articles, converted them to JSON documents, and cleaned up citations and bibliography links.

This section describes (Fig. 1) how we harnessed this dataset in order to *(1)* draw meaningful links between the articles of the CORD-19 corpus and the Web of Data by means of NEs, and *(2)* extract a graph of argumentative components discovered in the articles, while respecting the Semantic Web standards. The result of this work is referred to as the *Covid-on-the-Web dataset*.

2.1 Building the CORD-19 Named Entities Knowledge Graph

The *CORD-19 Named Entities Knowledge Graph* (CORD19-NEKG), part of the Covid-on-the-Web dataset, describes NEs identified and disambiguated in the articles of the CORD-19 corpus using three tools:

- DBpedia Spotlight [6] can annotate text in eight different languages with DBpedia entities. Disambiguation is carried out by entity linking using a generative model with maximum likelihood.
- Entity-fishing[6] identifies and disambiguates NEs against Wikipedia and Wikidata at document level. It relies on FastText word embeddings to generate candidates and ranks them with gradient tree boosting and features derived from relations and context.
- NCBO BioPortal Annotator [12] annotates biomedical texts against vocabularies loaded in BioPortal. Patterns are identified using the Mgrep method. Annotator+ [19] extends its capabilities with the integration of a lemmatizer and the Context/NegEx algorithms.

To ensure reusability, CORD19-NEKG leverages well-known, relevant terminological resources to represent articles and NEs in RDF. Below, we outline the

[6] `https://github.com/kermitt2/entity-fishing`

```
@prefix covidpr: <http://ns.inria.fr/covid19/property/>.
@prefix dct:     <http://purl.org/dc/terms/>.
@prefix oa:      <http://www.w3.org/ns/oa#>.
@prefix schema:  <http://schema.org/>.

[] a                 oa:Annotation;
   schema:about       <http://ns.inria.fr/covid19/f74923b3ce82c...>;
   dct:subject        "Engineering", "Biology";
   covidpr:confidence "1"^^xsd:decimal;
   oa:hasBody         <http://wikidata.org/entity/Q176996>;
   oa:hasTarget [
     oa:hasSource     <http://ns.inria.fr/covid19/f74923b3ce82c...#abstract>;
     oa:hasSelector   [ a oa:TextPositionSelector, oa:TextQuoteSelector;
       oa:exact "PCR"; oa:start "235"; oa:end "238" ]];
```

Listing 1.1. Representation of the "polymerase chain reaction" (PCR) named entity as an annotation of an article's abstract from offset 235 to 238.

main concepts of this RDF modeling. More details and examples are available on the project's Github repository.[7]

Article metadata (e.g., title, authors, DOI) and content are described using DCMI[8], Bibliographic Ontology (FaBiO)[9], Bibliographic Ontology[10], FOAF[11] and Schema.org[12]. NEs are modelled as annotations represented using the Web Annotation Vocabulary[13]. An example of annotation is given in Listing 1.1. The annotation body is the URI of the resource (e.g., from Wikidata) linked to the NE. The piece of text recognized as the NE itself is the annotation target. It points to the article part wherein the NE was recognized (title, abstract or body), and locates it with start and end offsets. Provenance information is also provided for each annotation (not shown in Listing 1.1) using PROV-O[14], that denotes the source being processed, the tool used to extract the NE, the confidence of extracting and linking the NE, and the annotation author.

2.2 Mining CORD-19 to Build an Argumentative Knowledge Graph

The **A**rgumentative **C**linical **T**rial **A**nalysis (ACTA) [13] is a tool designed to analyse clinical trials for argumentative components and PICO[15] elements. Originally developed as an interactive visualization tool to ease the work of clinicians in analyzing clinical trials, we re-purposed it to annotate the CORD-19 corpus. It goes far beyond basic keyword-based search by retrieving the main claim(s)

[7] https://github.com/Wimmics/covidontheweb/dataset

[8] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[9] https://sparontologies.github.io/fabio/current/fabio.html

[10] http://bibliontology.com/specification.html

[11] http://xmlns.com/foaf/spec/

[12] https://schema.org/

[13] https://www.w3.org/TR/annotation-vocab/

[14] https://www.w3.org/TR/prov-o/

[15] PICO is a framework to answer health-care questions in evidence-based practice that comprises patients/population (P), intervention (I), control/comparison (C) and outcome (O).

stated in the trial, as well as the evidence linked to this claim, and the PICO elements. In the context of clinical trials, a *claim* is a concluding statement made by the author about the outcome of the study. It generally describes the relation of a new treatment (intervention arm) with respect to existing treatments (control arm). Accordingly, an observation or measurement is an *evidence* which supports or attacks another argument component. Observations comprise side effects and the measured outcome. Two types of relations can hold between argumentative components. The *attack* relation holds when one component is contradicting the proposition of the target component, or stating that the observed effects are not statistically significant. The *support* relation holds for all statements or observations justifying the proposition of the target component.

Each abstract of the CORD-19 corpus was analyzed by ACTA and translated into RDF to yield the *CORD-19 Argumentative Knowledge Graph*. The pipeline consists of four steps: *(i)* the detection of argumentative components, i.e. claims and evidence, *(ii)* the prediction of relations holding between these components, *(iii)* the extraction of PICO elements, and *(iv)* the production of the RDF representation of the arguments and PICO elements.

Component Detection. This is a sequence tagging task where, for each word, the model predicts if the word is part of a component or not. We combine the BERT architecture[16] [7] with an LSTM and a Conditional Random Field to do token level classification. The weights in BERT are initialized with specialised weights from SciBERT [1] and provides an improved representation of the language used in scientific documents such as in CORD-19. The pre-trained model is fine-tuned on a dataset annotated with argumentative components resulting in .90 f_1-score [14]. As a final step, the components are extracted from the label sequences.

Relation Classification. Determining which relations hold between the components is treated as a three-class sequence classification problem, where the sequence consists of a pair of components, and the task is to learn the relation between them, i.e. *support, attack* or *no relation*. The SciBERT transformer is used to create the numerical representation of the input text, and combined with a linear layer to classify the relation. The model is fine-tuned on a dataset for argumentative relations in the medical domain resulting in .68 f_1-score [14].

PICO Element Detection. We employ the same architecture as for the component detection. The model is trained on the EBM-NLP corpus [17] to jointly predict the participant, intervention[17] and outcome candidates for a given input. Here, the f_1-score on the test set is .734 [13]. Each argumentative component is annotated with the PICO elements it contains. To facilitate structured queries, PICO elements are linked to Unified Medical Language System (UMLS) concepts with ScispaCy [16].

Argumentative knowledge graph. The *CORD-19 Argumentative Knowledge Graph* (CORD19-AKG) draws on the Argument Model Ontology (AMO)[18],

[16] BERT is a self-attentive transformer models that uses language model (LM) pre-training to learn a task-independent understanding from vast amounts of text in an unsupervised fashion.

[17] The intervention and comparison label are treated as one joint class.

[18] `http://purl.org/spar/amo/`

```
@prefix prov:   <http://www.w3.org/ns/prov#>.
@prefix schema:  <http://schema.org/>.
@prefix aif:    <http://www.arg.dundee.ac.uk/aif#>.
@prefix amo:    <http://purl.org/spar/amo/>.
@prefix sioca:  <http://rdfs.org/sioc/argument#>.

<http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496...>
  a               amo:Argument;
  schema:about    <http://ns.inria.fr/covid19/4f8d24c531d2c33496...>;
  amo:hasEvidence <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../0>;
  amo:hasClaim    <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../6>.

<http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../0>
  a amo:Evidence, sioca:Justification, aif:I-node;
  prov:wasQuotedFrom <http://ns.inria.fr/covid19/4f8d24c531d2c33496...>;
  aif:formDescription "17 patients discharged in recovered condition...";
  sioca:supports  <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../6>;
  amo:proves      <http://ns.inria.fr/covid19/arg/4f8d24c531d2c33496.../6>.
```

Listing 1.2. Example representation of argumentative components and their relation.

the SIOC Argumentation Module (SIOCA)[19] and the Argument Interchange Format[20]. Each argument identified by ACTA is modelled as an `amo:Argument` to which argumentative components (claims and evidence) are connected. The claims and evidences are themselves connected by support or attack relations (`sioca:supports`/`amo:proves` and `sioca:challenges` properties respectively). Listing 1.2 sketches an example. Furthermore, the PICO elements are described as annotations of the argumentative components wherein they were identified, in a way very similar to the NEs (as exemplified in Listing 1.1). Annotation bodies are the UMLS concept identifiers (CUI) and semantic type identifiers (TUI).

2.3 Automated Dataset Generation Pipeline

From a technical perspective, the CORD-19 corpus essentially consists of one JSON document per scientific article. Consequently, yielding the Covid-on-the-Web RDF dataset involves two main steps: process each document of the corpus to extract the NEs and arguments, and translate the output of both treatments into a unified, consistent RDF dataset. The whole pipeline is sketched in Fig. 1.

Named entities extraction. The extraction of NEs with DBpedia Spotlight, Entity-fishing and BioPortal Annotator produced approximately 150,000 JSON documents ranging from 100KB to 50 MB each. These documents were loaded into a MongoDB database, and pre-processed to filter out unneeded or invalid data (e.g., invalid characters) as well as to remove NEs that are less than three characters long. Then, each document was translated into the RDF model described in Section 2.1 using Morph-xR2RML,[21] an implementation of the xR2RML mapping language [15] for MongoDB databases. The three NE

[19] `http://rdfs.org/sioc/argument#`

[20] `http://www.arg.dundee.ac.uk/aif#`

[21] `https://github.com/frmichel/morph-xr2rml/`

extractors were deployed on a Precision Tower 5810 equipped with a 3.7GHz CPU and 64GB RAM. We used Spotlight with a pre-trained model[22] and Annotator's online API[23] with the Annotator+ features to benefit from the whole set of ontologies in BioPortal. To keep the files generated by Annotator+ of a manageable size, we disabled the options negation, experiencer, temporality, display_links and display_context. We enabled the longest_only option, as well as the lemmatization option to improve detection capabilities. Processing the CORD-19 corpus with the NEs extractors took approximately three days. MongoDB and Morph-xR2RML were deployed on a separate machine equipped with 8 CPU cores and 48GB RAM. The full processing, i.e., spanning upload in MongoDB of the documents produced by the NE extractors, pre-processing and RDF files generation, took approximately three days.

Argumentative graph extraction. Only the abstracts longer than ten sub-word tokens[24] were processed by ACTA to ensure meaningful results. In total, almost 30,000 documents matched this criteria. ACTA was deployed on a 2.8GHz dual-Xeon node with 96GB RAM, and processing the articles took 14 hours. Like in the NEs extraction, the output JSON documents were loaded into MongoDB and translated to the RDF model described in Section 2.2 using Morph-xR2RML. The translation to RDF was carried out on the same machine as above, and took approximately 10 minutes.

3 Publishing and Querying Covid-on-the-Web Dataset

The Covid-on-the-Web dataset consists of two main RDF graphs, namely the *CORD-19 Named Entities Knowledge Graph* and the *CORD-19 Argumentative Knowledge Graph*. A third, transversal graph describes the metadata and content of the CORD-19 articles. Table 1 synthesizes the amount of data at stake in terms of JSON documents and RDF triples produced. Table 2 reports some statistics against the different vocabularies used.

Dataset Description. In line with common data publication best practices [8], we paid particular attention to the thorough description of the Covid-on-the-Web dataset itself. This notably comprises (1) licensing, authorship and provenance information described with DCAT[25], and (2) vocabularies, interlinking and access information described with VOID[26]. The interested reader may look up the dataset URI[27] to visualize this information.

Dataset Accessibility. The dataset is made available by means of a DOI-identified RDF dump downloadable from Zenodo, and a public SPARQL endpoint. All URIs can be dereferenced with content negotiation. A Github repository provides a comprehensive documentation (including licensing, modeling,

[22] https://sourceforge.net/projects/dbpedia-spotlight/files/2016-10/en/

[23] http://data.bioontology.org/documentation

[24] Inputs were tokenized with the BERT tokenizer, where one sub-word token has a length of one to three characters.

[25] https://www.w3.org/TR/vocab-dcat/

[26] https://www.w3.org/TR/void/

[27] Covid-on-the-Web dataset URI: http://ns.inria.fr/covid19/covidontheweb-1-1

Table 1. Statistics of the Covid-on-the-Web dataset.

Type of data	JSON data	Resources produced	RDF triples
Articles metadata and textual content	7.4 GB	n.a.	1.27 M
CORD-19 Named Entities Knowledge Graph			
NEs found by DBpedia Spotlight (titles, abstracts)	35 GB	1.79 M	28.6 M
NEs found by Entity-fishing (titles, abstracts, bodies)	23 GB	30.8 M	588 M
NEs found by BioPortal Annotator (titles, abstracts)	17 GB	21.8 M	52.8 M
CORD-19 Argumentative Knowledge Graph			
Claims/evidence components (abstracts)	138 MB	53 K	545 K
PICO elements		229 K	2.56 M
Total for Covid-on-the-Web (including articles metadata and content)			
	82 GB	54 M named entities 53 K claims/evidence 229 K PICO elements	674 M

Table 2. Selected statistics on properties/classes/resources.

Property URI	nb of instances	comments
http://purl.org/dc/terms/subject	62,922,079	Dublin Core subject property
http://www.w3.org/ns/oa#hasBody	54,760,308	Annotation Ontology body property
http://schema.org/about	34,971,108	Schema.org about property
http://www.w3.org/ns/prov#wasGeneratedBy	34,971,108	PROV-O "generated by" property
http://purl.org/dc/terms/creator	34,741,696	Dublin Core terms creator property
http://purl.org/spar/cito/isCitedBy	207,212	CITO citation links
http://purl.org/vocab/frbr/core#partOf	114,021	FRBR part of relations
http://xmlns.com/foaf/0.1/surname	65,925	FOAF surnames

Class URI	nb of instances	comments
http://www.w3.org/ns/oa#Annotation	34,950,985	Annotations from the Annotation Ontology
http://www.w3.org/ns/prov#Entity	34,721,578	Entities of PROV-O
http://purl.org/spar/amo/Claim	28,140	Claims from AMO
http://rdfs.org/sioc/argument#Justification	25,731	Justifications from SIOC Argument

Resource URI	nb of uses	comments
http://www.wikidata.org/entity/Q103177	209,183	severe acute respiratory syndrome (Wikidata)
http://purl.obolibrary.org/obo/NCBITaxon_10239	11,488	Virus in NCBI organismal classification
http://dbpedia.org/resource/Severe_acute_respiratory_syndrome	5,280	severe acute respiratory syndrome (DBpedia)
http://www.ebi.ac.uk/efo/EFO_0005741	8,753	Infectious disease in Experimental Factor Ontology

named graphs and third-party vocabularies loaded in the SPARQL endpoint). This information is summarized in Table 3.

Reproducibility. In compliance with the open science principles, all the scripts, configuration and mapping files involved in the pipeline are provided in the project's Github repository under the terms of the Apache License 2.0, so that anyone may rerun the whole processing pipeline (from articles mining to loading RDF files into Virtuoso OS).

Dataset Licensing. Being derived from the CORD-19 dataset, different licences apply to the different subsets of the Covid-on-the-Web dataset. The subset corresponding to the CORD-19 dataset translated into RDF (including articles metadata and textual content) is published under the terms of the CORD-19

Table 3. Dataset availability.

Dataset DOI	10.5281/zenodo.3833753
Downloadable RDF dump	`https://doi.org/10.5281/zenodo.3833753`
Public SPARQL endpoint	`https://covidontheweb.inria.fr/sparql`
Documentation	`https://github.com/Wimmics/CovidOnTheWeb`
URIs namespace	`http://ns.inria.fr/covid19/`
Dataset URI	`http://ns.inria.fr/covid19/covidontheweb-1-1`
Citation	Wimmics Research Team. (2020). Covid-on-the-Web dataset (Version 1.1). Zenodo. `http://doi.org/10.5281/zenodo.3833753`

license.[28] In particular, this license respects the sources that are copyrighted. The subset produced by mining the articles, either the NEs (CORD19-NEKG) or argumentative components (CORD19-AKG) is published under the terms of the Open Data Commons Attribution License 1.0 (ODC-By).[29]

Sustainability Plan. In today's context, where new research is published weekly about the COVID-19 topic, the value of Covid-on-the-Web, as well as other related datasets, lies in the ability to keep up with the latest advances and ingest new data as it is being published. Towards this goal, we've taken care of producing a documented, repeatable pipeline, and we have already performed such an update thus validating the procedure. In the middle-term, we intend to improve the update frequency while considering *(1)* the improvements delivered by CORD-19 updates, and *(2)* the changing needs to be addressed based on the expression of new application scenarios (see Section 5). Furthermore, we have deployed a server to host the SPARQL endpoint that benefits from a high-availability infrastructure and 24/7 support.

4 Visualization and Current Usage of the Dataset

Beyond the production of the Covid-on-the-Web dataset, our project has set out to explore ways of visualizing and interacting with the data. We have developed a tool named *Covid Linked Data Visualizer*[30] comprising a query web interface hosted by a node.js server, a transformation engine based on the Corese Semantic Web factory [5], and the MGExplorer graphic library [4]. The web interface enables users to load predefined SPARQL queries or edit their own queries, and execute them against our public SPARQL endpoint. The queries are parameterized by HTML forms by means of which the user can specify search criterion, e.g., the publication date. The transformation engine converts the JSON-based SPARQL results into the JSON format expected by the graphic library. Then, exploration of the result graph is supported by MGExplorer that encompasses a set of specialized visualization techniques, each of them allowing to focus on a

[28] CORD-19 license `https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/`

[29] ODC-By license: `http://opendatacommons.org/licenses/by/1.0/`

[30] Covid Linked Data Visualizer can be tested at: `http://covid19.i3s.unice.fr:8080`

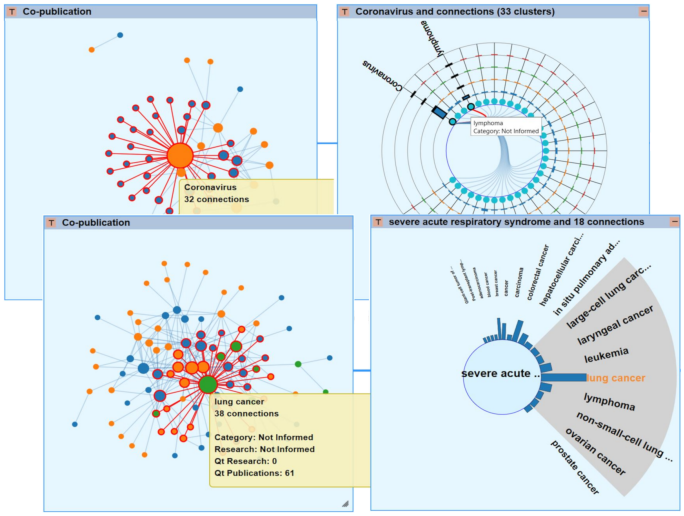

Fig. 2. Covid Linked Data Visualizer: visualization of the subset of articles that mention both a type of cancer (blue dots) and a virus of the corona family (orange dots).

particular type of relationship. Fig. 2 illustrates some of these techniques: node-edge diagram (left) shows an overview of all the nodes and their relationships; ClusterVis (top right) is a cluster-based visualization allowing the comparison of node attributes while keeping the representation of the relationships among them; IRIS (bottom right) is an egocentric view for displaying all attributes and relations of a particular node. The proposed use of information visualization techniques is original in that it provides users with interaction modes that can help them explore, classify and analyse the importance of publications. This is a key point for making the tools usable and accessible, and get adoption.

During a meeting with some health and medical research organisations (i.e., Inserm and INCa), an expert provided us with an example query that researchers would be interested in solving against a dataset like the one we generated: "find the articles that mention both a type of cancer and a virus of the corona family". Taking that query as a first competency question, we used Covid Linked Data Visualizer whose results are visualized with the MGExplorer library (Fig. 2). We also created several Python and R Jupyter notebooks[31] to demonstrate the transformation of the result into structures such as Dataframes[32] for further analysis (Fig. 3).

Let us finally mention that, beyond our own uses, the Covid-on-the-Web dataset is now served by the LOD Cloud cache hosted by OpenLink Software.[33]

[31] https://github.com/Wimmics/covidontheweb/tree/master/notebooks

[32] Dataframes are tabular data structures widely used in Python and R for the data analysis.

[33] https://twitter.com/kidehen/status/1250530568955138048

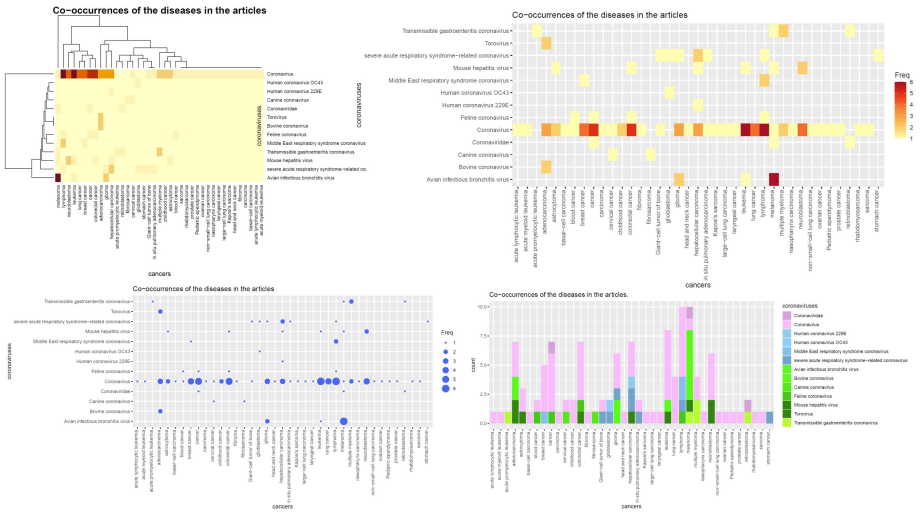

Fig. 3. Visualizations of Jupyter Notebook query results: four different representations of the number of articles that co-mention cancer types and viruses of corona family.

5 Potential Impact and Reusability

To the best of our knowledge, the Covid-on-the-Web dataset is the first one integrating NEs, arguments and PICO elements into a single, coherent whole. We are confident that it will serve as a foundation for Semantic Web applications as well as for benchmarking algorithms and will be used in challenges. The resources and services that we offer on the COVID-19 literature are of interest for health organisations and institutions to extract and intelligently analyse information on a disease which is still relatively unknown and for which research is constantly evolving. To a certain extent, it is possible to cross-reference information to have a better understanding of this matter and, in particular, to initiate research into unexplored paths. We also hope that the openness of the data and code will allow contributors to advance the current state of knowledge on this disease which is impacting the worldwide society. In addition to being interoperable with central knowledge graphs used within the Semantic Web community, the visualizations we offer through MGExplorer and Notebooks show the potential of these technologies in other fields, e.g., the biomedical and medical ones.

Interest of communities in using the Dataset and Services. Several biomedical institutions have shown interest in using our resources, eithet direct project partners (French Institute of Medical Research - Inserm, French National Cancer Institute - INCa) or indirect (Antibes Hospital, Nice Hospital). For now, these institutions act as potential users of the resources, and as co-designers. Furthermore, given the importance of the issues at stake and the strong support that they can provide in dealing with them, we believe that other similar institutions could be interested in using the resources.

Documentation/Tutorials. For design rationale purposes, we keep records of the methodological documents we use during the design of the resources (e.g., query elicitation documents), the technical documentation of the algorithms and models[34], the best practices we follow (FAIR, Cool URIs, five-star linked data, etc.) and the end users help (e.g., demonstration notebooks).

Application scenarios, user models, and typical queries. Our resources are based on generic tools that we are adapting to the COVID-19 issue. Precisely, having a user-oriented approach, we are designing them according to three main motivating scenarios identified through a need analysis of the biomedical institutions with whom we collaborate:
Scenario 1: Helping clinicians to get argumentative graphs to analyze clinical trials and make evidence-based decisions.
Scenario 2: Helping hospital physicians to collect ranges of human organism's substances (e.g., cholesterol) from scientific articles, to determine if the substances' levels of their patients are normal or not.
Scenario 3: Helping missions heads from a Cancer Institute to collect scientific articles about cancer and coronavirus to elaborate research programs to deeper study the link between cancer and coronavirus.

The genericity of the basic tools will allow us to later on apply the resources to a wider set of scenarios, and our biomedical partners already urge us to start thinking of scenarios related to other issues than the COVID-19.

Besides the scenarios above, we are also eliciting representative user models (in the form of personas), the aim of which is to help us – as service designers – to understand our users' needs, experiences, behaviors and goals.

We also elicited meaningful queries from the potential users we interviewed. These queries serve to specify and test our knowledge graph and services. For genericity purposes, we elaborated a typology from the collected queries, using dimensions such as: Prospective vs. Retrospective queries or Descriptive (requests for description) vs. Explanatory (requests for explanation) vs. Argumentative (requests for argumentation) queries. Here are examples of such queries:
- *Prospective descriptive queries:* What types of cancers are likely to occur in COVID-19 victims in the next years? In what types of patients? Etc.
- *Descriptive retrospective queries:* What types of cancers appeared in [SARSCoV1 | MERS-CoV] victims in the [2|3|n] years that followed? What was the rate of occurrence? In what types of patients? Etc. What are the different sequelae related to Coronaviruses? Which patients cured of COVID-19 have pulmonary brosis?
- *Retrospective explanatory queries:* Did [SARS-CoV1 | MERS-CoV] cause cancer? Was [cell transformation | cancer development] caused directly by coronavirus infection? Or was it caused indirectly through [inflammation | metabolic changes] caused by this infection? Which coronavirus-related sequelae are responsible for the greatest potential for cell transformation?
- *Argumentative retrospective queries:* What is the evidence that [SARSCoV1 | MERS-CoV] caused cancer? What experiments have shown that the pulmonary brosis seen in patients cured of COVID-19 was caused by COVID-19?

[34] https://github.com/Wimmics/covidontheweb/blob/master/doc/01-data-modeling.md

These queries are a brief illustration of an actual (yet non-exhaustive) list of questions raised by users. It is worthy of notice that whilst some questions might be answered by showing the correlation between components (e.g., types of cancer), others require the representation of trends (e.g., cancer likely to occur in the next years), and analysis of specific attributes (e.g., details about metabolic changes caused by COVID-19). Answering these complex queries requires exploration of the CORD-19 corpus, and for that we offer a variety of analysis and visualization tools. These queries and the generic typology shall be reused in further extensions and other projects.

The Covid Linked Data Visualizer (presented in section 4) supports the visual exploration of the Covid-on-the-Web dataset. Users can inspect the attributes of elements in the graph resulting from a query (by positioning the mouse over elements) or launch a chained visualization using any of the interaction techniques available (ex. IRIS, ClusterVis, etc). These visualization techniques are meant to help users understand the relationships available in the results. For example, users can run a query to visualize a co-authorship network; then use IRIS and ClusterVis to understand who is working together and on which topics. They can also run a query looking for papers mentioning the COVID-19 and diverse types of cancer. Finally, the advanced mode makes it possible to add new SPARQL queries implementing other data exploration chains.

6 Related Works

Since the first release of the CORD-19 corpus, multiple initiatives, ranging from quick-and-dirty data releases to the repurposing of existing large projects, have started analyzing and mining it with different tools and for different purposes. Entity linking is usually the first step to further processing or enriching. Hence, not surprisingly, several initiatives have already applied these techniques to the CORD-19 corpus. **CORD-19-on-FHIR**[35] results of the translation of the CORD-19 corpus in RDF following the HL7-FHIR interchange format, and the annotation of articles with concepts related to conditions, medications and procedures. The authors also used Pubtator [21] to further enrich the corpus with concepts such as gene, disease, chemical, species, mutation and cell line. **KG-COVID-19**[36] seeks the lightweight construction of KGs for COVID-19 drug repurposing efforts. The KG is built by processing the CORD-19 corpus and adding NEs extracted from COVIDScholar.org and mapped to terms from biomedical ontologies. **Covid9-PubAnnotation**[37] is a repository of text annotations concerning CORD-19 as well as LitCovid and others. Annotations are aggregated from multiple sources and aligned to the canonical text that is taken from PubMed and PMC. The **Machine Reading for COVID-19 and Alzheimer's**[38] project aims at producing a KG representing causal inference

[35] https://github.com/fhircat/CORD-19-on-FHIR

[36] https://github.com/Knowledge-Graph-Hub/kg-covid-19/

[37] https://covid19.pubannotation.org/

[38] https://github.com/kingfish777/COVID19

extracted from semantic relationships between entities such as drugs, biomarkers or comorbidities. The relationships were extracted from the Semantic MEDLINE database enriched with CORD-19. **CKG-COVID-19**[39] seeks the discovery of drug repurposing hypothesis through link prediction. It processed the CORD-19 corpus with state of the art machine reading systems to build a KG where entities such as genes, proteins, drugs, diseases, etc. are linked to their Wikidata counterparts.

When comparing Covid-on-the-Web with these other initiatives, several main differences can be pointed out. First, they restrict processing to the title and abstract of the articles, whereas we process the full text of the articles with Entityfishing, thus providing a high number of NEs linked to Wikidata concepts. Second, these initiatives mostly focus on biomedical ontologies. As a result, the NEs identified typically pertain to genes, proteins, drugs, diseases, phenotypes and publications. In our approach, we have not only considered biomedical ontologies from BioPortal, but we have also extended this scope with two general knowledge bases that are major hubs in the Web of Data: DBpedia and Wikidata. Finally, to the best our knowledge, our approach is the only one to integrate argumentation structures and named entities in a coherent dataset.

Argument(ation) Mining (AM) [3] is the research area aiming at extracting and classifying argumentative structures from text. AM methods have been applied to heterogeneous types of textual documents. However, only few approaches [22,11,14] focused on automatically detecting argumentative structures from textual documents in the medical domain, e.g., clinical trials, guidelines, Electronic Health Records. Recently, transformer-based contextualized word embeddings have been applied to AM tasks [18,14]. To the best of our knowledge, Covid-on-the-Web is the first attempt to apply AM to the COVID-19 literature.

7 Conclusion and Future Works

In this paper, we described the data and software resources provided by the Covid-on-the-Web project. We adapted and combined tools to process, analyze and enrich the CORD-19 corpus, to make it easier for biomedical researchers to access, query and make sense of COVID-19 related literature. We designed and published a linked data knowledge graph describing the named entities mentioned in the CORD-19 articles and the argumentative graphs they include. We also published the pipeline we set up to generate this knowledge graph, in order to *(1)* continue enriching it and *(2)* spur and facilitate reuse and adaptation of both the dataset and the pipeline. On top of this knowledge graph, we developed, adapted and deployed several tools providing Linked Data visualizations, exploration methods and notebooks for data scientists. Through active interactions (interviews, observations, user tests) with institutes in healthcare and medical research, we are ensuring that our approach is guided by and aligned with the actual needs of the biomedical community. We have shown that with our dataset, we can perform documentary research and provide visualizations

[39] https://github.com/usc-isi-i2/CKG-COVID-19

suited to the needs of experts. Great care has been taken to produce datasets and software that meet the open and reproducible science goals and the FAIR principles.

We identified that, since the emergence of the COVID-19, the unusual pace at which new research has been published and knowledge bases have evolved raises critical challenges. For instance, a new release of CORD-19 is published weekly, which challenges the ability to keep up with the latest advances. Also, the extraction and disambiguation of NEs was achieved with pre-trained models produced before the pandemic, typically before the SARS-Cov-2 entity was even created in Wikidata. Similarly, it is likely that existing terminological resources are being/will be released soon with COVID-19 related updates. Therefore, in the middle term, we intend to engage in a sustainability plan aiming to routinely ingest new data and monitor knowledge base evolution so as to reuse updated models. Furthermore, since there is no reference CORD-19 subset that has been manually annotated and could serve as ground truth, it is hardy possible to evaluate the quality of the machine learning models used to extract named entities and argumentative structures. To address this issue, we are currently working on the implementation of data curation techniques, and the automated discovery of frequent patterns and association rules that could be used to detect mistakes in the extraction of named entities, thus allowing to come up with quality enforcing measures.

References

1. I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.
2. M. Bersanelli. Controversies about COVID-19 and anticancer treatment with immune checkpoint inhibitors. *Immunotherapy*, 12(5):269–273, Apr. 2020.
3. E. Cabrio and S. Villata. Five years of argument mining: a data-driven analysis. In *Proc. of IJCAI 2018*, pages 5427–5433, 2018.
4. R. A. Cava, C. M. D. S. Freitas, and M. Winckler. Clustervis: visualizing nodes attributes in multivariate graphs. In A. Seffah, B. Penzenstadler, C. Alves, and X. Peng, editors, *Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, April 3-7, 2017*, pages 174–179. ACM, 2017.
5. O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the semantic web with Corese search engine. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, volume 16, page 705, Valencia, Spain, 2004.
6. J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124, 2013.
7. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.
8. B. Farias Lóscio, C. Burle, and N. Calegari. Data on the Web Best Practices. *W3C Recommandation*, 2017.

9. R. Gazzotti, C. Faron-Zucker, F. Gandon, V. Lacroix-Hugues, and D. Darmon. Injecting domain knowledge in electronic medical records to improve hospitalization prediction. In P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. G. Gray, V. López, A. Haller, and K. Hammar, editors, *The Semantic Web - 16th European Conference, ESWC, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503 of *Lecture Notes in Computer Science*, pages 116–130. Springer, 2019.
10. R. Gazzotti, C. Faron-Zucker, F. Gandon, V. Lacroix-Hugues, and D. Darmon. Injection of automatically selected DBpedia subjects in electronic medical records to boost hospitalization prediction. In C. Hung, T. Cerný, D. Shin, and A. Bechini, editors, *SAC '20: The 35th ACM/SIGAPP Symposium on Applied Computing, online event, March 30 - April 3, 2020*, pages 2013–2020. ACM, 2020.
11. N. Green. Argumentation for scientific claims in a biomedical research article. In *Proc. of ArgNLP 2014 workshop*, 2014.
12. C. Jonquet, N. H. Shah, and M. A. Musen. The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56, 2009.
13. T. Mayer, E. Cabrio, and S. Villata. ACTA a tool for argumentative clinical trial analysis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6551–6553, 2019.
14. T. Mayer, E. Cabrio, and S. Villata. Transformer-based argument mining for healthcare applications. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020.
15. F. Michel, L. Djimenou, C. Faron-Zucker, and J. Montagnat. Translation of Relational and Non-Relational Databases into RDF with xR2RML. In *Proceeding of the 11th International Conference on Web Information Systems and Technologies (WebIST)*, pages 443–454, Lisbon, Portugal, 2015.
16. M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, Aug. 2019. Association for Computational Linguistics.
17. B. Nye, J. J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, and B. Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 197–207, 2018.
18. N. Reimers, B. Schiller, T. Beck, J. Daxenberger, C. Stab, and I. Gurevych. Classification and clustering of arguments with contextualized word embeddings. In *Proc. of ACL 2019*, pages 567–578, 2019.
19. A. Tchechmedjiev, A. Abdaoui, V. Emonet, S. Melzi, J. Jonnagaddala, and C. Jonquet. Enhanced functionalities for annotating and indexing clinical text with the ncbo annotator+. *Bioinformatics*, 34(11):1962–1965, 2018.
20. L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. M. Kinney, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. Cord-19: The covid-19 open research dataset. *ArXiv*, abs/2004.10706, 2020.
21. C.-H. Wei, H.-Y. Kao, and Z. Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522, 2013.
22. J. Zabkar, M. Mozina, J. Videcnik, and I. Bratko. Argument based machine learning in a medical domain. In *Proc. of COMMA 2006*, pages 59–70, 2006.