



RESEARCH ARTICLE

Covid-on-the-Web: Exploring the COVID-19 scientific literature through visualization of linked data from entity and argument mining

Aline Menin¹, Franck Michel¹, Fabien Gandon¹, Raphaël Gazzotti¹, Elena Cabrio¹, Olivier Corby¹, Alain Giboin¹, Santiago Marro¹, Tobias Mayer¹, Serena Villata¹, and Marco Winckler¹

University Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France

an open access journal



Citation: Menin, A., Michel, F., Gandon, F., Gazzotti, R., Cabrio, E., ... Winckler, M. (2021). Covid-on-the-Web: Exploring the COVID-19 scientific literature through visualization of linked data from entity and argument mining. *Quantitative Science Studies*, 2(4), 1301–1323. https://doi.org/10.1162/qss_a_00164

DOI: https://doi.org/10.1162/qss_a_00164

Corresponding Author:
Aline Menin
aline.menin@inria.fr

Keywords: argument mining, COVID-19, entity linking, linked data, visualization

ABSTRACT

The unprecedented mobilization of scientists caused by the COVID-19 pandemic has generated an enormous number of scholarly articles that are impossible for a human being to keep track of and explore without appropriate tool support. In this context, we created the Covid-on-the-Web project, which aims to assist the accessing, querying, and sense-making of COVID-19-related literature by combining efforts from the semantic web, natural language processing, and visualization fields. In particular, in this paper we present an RDF data set (a linked version of the “COVID-19 Open Research Dataset” (CORD-19), enriched via entity linking and argument mining) and the “Linked Data Visualizer” (LDViz), which assists the querying and visual exploration of the referred data set. The LDViz tool assists in the exploration of different views of the data by combining a querying management interface, which enables the definition of meaningful subsets of data through SPARQL queries, and a visualization interface based on a set of six visualization techniques integrated in a chained visualization concept, which also supports the tracking of provenance information. We demonstrate the potential of our approach to assist biomedical researchers in solving domain-related tasks, as well as to perform exploratory analyses through use case scenarios.

1. INTRODUCTION

The COVID-19 pandemic has motivated the scientific community in numerous fields of research to contribute in a common effort to study, understand, and fight the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Several data sets covering the publications about COVID-19 and related coronaviruses and diseases have been compiled to support the scientific community. In particular, we focus on the COVID-19 Open Research Dataset (CORD-19) (Wang, Lo et al., 2020), which gathers over 500,000 scholarly articles, including over 200,000 with full text. This deluge of ever-increasing publications in such a short time frame suggests that it is impossible for any researcher to examine every publication and extract the relevant information from it without appropriate support. To help researchers find publications of interest, we employ information visualization techniques to explore the data set and identify relationships among publications that indicate those that are worthy of further examination.

Copyright: © 2021 Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, and Marco Winckler. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.



In collaboration with biomedical researchers from the French Institute of Medical Research (Inserm)¹ and the French National Cancer Institute (INCa)², we created the Covid-on-the-Web project, which gathers expertise from various research fields (i.e., the semantic web, natural language processing, and visualization) to assist in the exploration of the COVID-19 scientific literature. Through a series of interviews with our prospective users, we could identify a set of meaningful use case scenarios, such as determining the right amount of certain substances in the patients' organism using baseline information collected from scientific articles, analyzing clinical trials to make evidence-based decisions, studying the relationship between coronaviruses and other diseases (e.g., cancer), and identifying the types of cancer that are likely to occur in COVID-19 victims, among others. Although some scenarios require exploring the relationship between components (e.g., cancer and coronavirus), others require representing trends (e.g., probability of cancer in COVID-19 victims) and analyzing specific attributes (e.g., details about metabolic changes caused by COVID-19). Furthermore, the analysis of coauthorship is relevant to health research as it allows us to assess collaboration trends and identify leading investigators and organizations (Fonseca, Sampaio et al., 2016). In this paper, we focus on using visualization to assist the resolution of user queries based on the relationship between components and coauthorship networks, which allow us to answer user queries such as "Where is research in a particular topic being performed?"

We present two contributions of the Covid-on-the-Web project to the exploration of COVID-19 scientific literature. The first contribution refers to the Covid-on-the-Web RDF data set, a linked version of the COVID-19 corpus, enriched via entity linking and argument mining. Currently, the Covid-on-the-Web RDF data set includes and enriches over 100,000 full-text scholarly articles from the 47th version of the COVID-19 corpus, which corresponds to 1.3 billion RDF triples describing the articles' metadata, an argumentation, and a named entities (NE) knowledge graph. The second contribution corresponds to LDViz³, a visualization tool that enables the exploration of the COVID-19 scientific literature from different perspectives, such as coauthorship, NE co-occurrence and the relationship between claims and evidence within publications. We demonstrate the potential of LDViz to support the exploration of customizable SPARQL result sets extracted from the Covid-on-the-Web data set to assist the resolution of different domain-related tasks.

Although there have been previous contributions in exploring the COVID-19 corpus through entity linking approaches (e.g., Oniani, Jiang et al., 2020; Reese, Unni et al., 2021), to the best of our knowledge, the Covid-on-the-Web data set is the first to integrate NE, arguments and PICO components into a single, coherent whole. Furthermore, we propose a unified pipeline (Figure 1) that facilitates the extraction and visualization of information from the COVID-19 corpus by continuously producing and publishing an enriched linked data knowledge graph. Also, our visualization approach differs from previous solutions to exploring the COVID-19 scientific literature (e.g., Hope, Portenoy et al., 2020; Verspoor, Šuster et al., 2020), by supporting the exploration of meaningful subsets of data suitable to users' needs through the definition of custom SPARQL SELECT queries and via multiple, complementary visualization techniques and by allowing the user to trace back their exploratory path, which helps them to understand how they have arrived at a certain outcome.

The remainder of this paper is organized as follows. Section 2 presents previous data mining and visualization approaches to exploring the COVID-19 corpus. Section 3 describes

¹ <https://www.inserm.fr/>

² <https://www.e-cancer.fr/>

³ Illustration video of LDViz: https://youtu.be/Cn_IWQ7yVvE

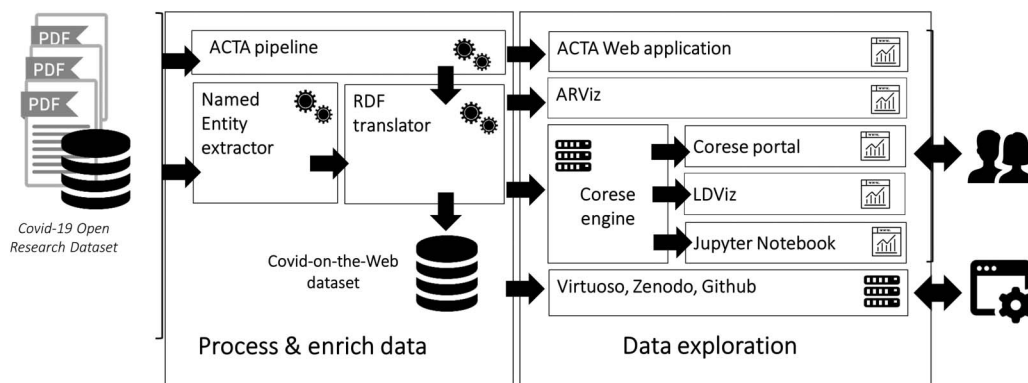


Figure 1. Overview of the Covid-on-the-Web project: Pipeline, resources, services and applications.

the extraction pipeline to process the COVID-19 corpus and generate the RDF data set and presents the characteristics of the data set and the available services to exploit it. Section 4 describes LDViz, which usage and exploration potentials are demonstrated through use case scenarios in Section 5. Section 6 discusses future applications and potential impact of the data set. Finally, Section 7 concludes the paper.

2. RELATED WORK

Since March 2020, when the COVID-19 corpus was first released, we have seen multiple efforts towards its analysis and mining through different tools and for various purposes. We have seen initiatives ranging from *ad hoc* data releases to the repurposing of large existing projects. Thus, in this section, we will present previous work related to the exploration of the COVID-19 data set in terms of data enrichment and visualization.

2.1. Data Enrichment

Entity linking is usually the first approach for processing or enriching a data set, which we can observe in several initiatives throughout the literature, such as the COVID-19-on-FHIR (Oniani et al., 2020) project, which transforms the COVID-19 corpus in RDF following the HL7-FHIR interchange format and annotates articles with concepts related to conditions, medications, and procedures; the KG-COVID-19 (Reese et al., 2021) project, which seeks the lightweight construction of KGs for COVID-19 drug repurposing efforts; and the CKG-COVID-19 (Ilievski, Garijo et al., 2020) project, which seeks the discovery of drug repurposing hypotheses through link prediction.

These solutions restrict processing to title and abstract, while we process the full text of the articles with Entity-fishing, thus providing a high number of NE linked to Wikidata concepts. Furthermore, these solutions are mostly focused on biomedical ontologies, resulting in NE strongly related to genes, proteins, drugs, diseases, phenotypes, and publications, while we extend the scope of ontologies to include DBpedia and Wikidata, resulting in NE that go beyond the biological domain to extend the scope of analysis. Furthermore, we integrate argumentation structures and NE in a coherent data set.

2.2. Visualization Approaches

The Covid19-PubAnnotation⁴ repository gathers text annotations regarding the COVID-19 corpus and other COVID-19 data sets. The annotations are recovered from multiple sources and

⁴ <https://covid19.pubannotation.org/>

aligned to the canonical text that is taken from PubMed and PMC archives, which link annotations to each other. Furthermore, the platform provides simple visualization that allows one to view the annotations directly on the text and further explore them through interaction.

The SciSight (Hope et al., 2020) tool enables exploratory search of the COVID-19 scientific literature and supports browsing through networks of biomedical concepts and research groups. It automatically extracts textual and coauthorship network information from publications, which are then explored through multiple views: A collocation explorer based on a non-ribbon chord diagram is used to represent the association between terms co-occurring in the same sentence; the relationship between patient characteristics and interventions (P and I from PICO elements) can be explored through two coordinated bar charts, which also display the temporal distribution of publications related to those criteria through a time series chart; and a network diagram represents the relationship between groups of coauthors defined either by social (shared authors) or topical affinity.

The COVID-SEE (Scientific Evidence Explorer for COVID-19) interface (Verspoor et al., 2020) enables the visual exploration of documents from the COVID-19 corpus through three different views: A Sankey diagram displays the relationship between PICO concepts and allows us to retrieve the documents where these relations occur; a topic view shows the representative topics of the selected documents and their distribution according to certain coherence measures; and a word cloud view displays the representative concepts of a document.

The SemViz (Tu, Verhagen et al., 2020) interface uses semantic visualization to explore the publications within the COVID-19 and other COVID-19 data sets. It provides three visualization techniques: A tag cloud gives an overall view of the most important concepts within the data; a heat map represents a pairwise relationship between selected entities in the article abstracts and journal names; and a data table is used to represent indexed document data, such as sentences of biomedical relations and corresponding PubMed URLs that link to the full article.

Sukla, Naskar et al. (2021) propose a visualization interface that allows the user to explore a set of publications from the COVID-19 corpus retrieved via textual querying. It displays the list of articles related to the query, from which corresponding NE can be further explored through a tag cloud chart and a co-occurrence map.

Bras, Gharavi et al. (2020) combine advanced data modeling of large corpora, information mapping, and trend analysis to provide a browsing and search interface for discovering topics and research resources within the COVID-19 data set. The system provides a cluster visualization displaying all resources in the data set, where the user can select a resource to explore its related topics, descriptions, trend analysis, and documents.

The CovidExplorer (Ambavi, Vaishnav et al., 2020) is a multifaceted AI-based search and visualization engine that integrates search and recommendation, statistics, and social media discussions to support the exploration of scientific articles from the COVID-19 data set. It comprises a query interface that supports keyword-based search of authors, papers (title), and full-text papers; and a named entity recognition system that computes indicators of first mention of entities, popular mentioned entities, and year-wise distribution of mention frequencies. These indicators are visualized through a timeline chart and a Sankey diagram, which shows the co-occurrence of entities within publications. The system provides a spatiotemporal visualization of tweets regarding COVID-19.

Although we find several visualization tools to support either the exploration of linked data in general or the COVID-19 scientific literature, such as the ones presented above, most of

them support the exploration of raw data (i.e., the RDF graph, OWL or RDF Schema), which is interesting for certain tasks, such as exploring the relevant concepts of an application domain via ontology representation, inspecting RDF Graphs, and analyzing instances based on their types/classes. Thus, we propose a flexible tool to enable users to define meaningful data sets via SPARQL SELECT queries applied to any SPARQL endpoint (illustrated here via the Covid-on-the-Web data set), so that they can explore multiple aspects of RDF data sets and the LOD Cloud. It also allows users to perform exploratory searches using various complementary visualization techniques instantiated on demand according to the task at hand, instead of a single visualization technique that represents the whole data set, restraining the analysis to a single view to the data. Our approach is also based on a visualization concept that enables users to track their exploratory path to help them to understand how they arrived to a certain outcome and to allow them to explore alternative hypotheses generated on the fly through different exploratory paths. Furthermore, the visualization together with the additional extractions (i.e., NEs, arguments) we perform in the Covid-on-the-Web data set, enables a deep and semantic-aware exploration of the topics and claims of the COVID scientific corpus by leveraging the combination of semantic processing and exploratory search.

3. THE “COVID-ON-THE-WEB” DATA SET

In this section, we describe the Covid-on-the-Web data set which we produced by processing and analyzing the CORD-19 corpus. The data set cohesively integrates the results of two mining processes: an NE extraction and linking that defines the links between the CORD-19 articles and major public data sets of the Web of Data, and an extraction of argumentative components discovered in the articles. These are both represented as RDF knowledge graphs described hereafter.

3.1. The CORD-19 Named Entities Knowledge Graph

The CORD-19 Named Entities Knowledge Graph (CORD19-NEKG) represents NE identified and disambiguated in the articles of the CORD-19 corpus using three tools: DBpedia Spotlight (Daiber, Jakob et al., 2013) to disambiguate NE against DBpedia entities; the Entity-fishing⁵ tool to disambiguate NE against Wikidata entities; and NCBO BioPortal Annotator (Jonquet, Shah et al., 2009) to disambiguate NE against entities found in BioPortal’s ontologies.

CORD19-NEKG uses common, well-adopted terminological resources to represent articles and NE in RDF. We use DCMI⁶, FaBio⁷, the Bibliographic Ontology⁸, FOAF⁹, and Schema.org¹⁰ to represent article metadata such as the title, authors, and DOI, and the Web Annotation Vocabulary¹¹ and Provenance Ontology¹² to represent and trace the recognized entities. These include the text segment recognized as the NE, the location of the segment within the article’s text, the resource URI (e.g., from Wikidata) linked to the NE, and the part of the article wherein the NE was recognized (i.e., title, abstract, or body). Figure 2 presents an extract of the RDF model, a full description of which, together with examples, is available in the project’s Github repository.¹³

⁵ <https://github.com/kermitt2/entity-fishing>

⁶ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

⁷ <https://sparontologies.github.io/fabio/current/fabio.html>

⁸ <https://bibliontology.com/specification.html>

⁹ <https://xmlns.com/foaf/spec/>

¹⁰ <https://schema.org/>

¹¹ <https://www.w3.org/TR/annotation-vocab/>

¹² <https://www.w3.org/TR/prov-o/>

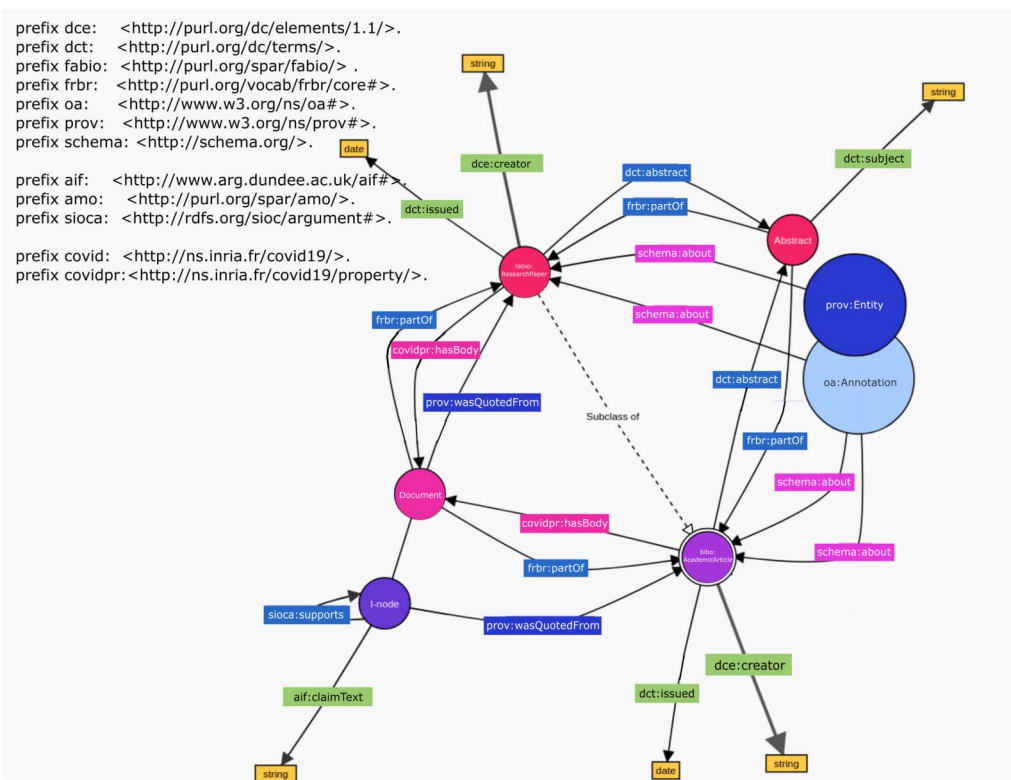


Figure 2. Extract of the Covid-on-the-Web RDF graph. Image adapted from an illustration generated with LD-VOWL (Lohmann, Negru et al., 2016) (see <https://vowl.visualdataweb.org/v2/> for a description of the graphical primitives and color scheme).

3.2. The COVID-19 Argumentative Knowledge Graph

The ACTA (Argumentative Clinical Trial Analysis) (Mayer, Cabrio, & Villata, 2019) tool was originally designed to help clinicians make decisions in evidence-based medicine by automatically extracting argumentative components and PICO elements¹⁴ from clinical trials. Through multiple NLP steps, ACTA retrieves the argumentative components in the trial and its PICO elements, classifies the components into *claim* (concluding statement) and *evidence* (observation or measurement), and infers the relationship between the components (i.e., *support* or *attack*). For instance, “a new treatment is considered more effective than existing treatments (claim), as attested by the measure of certain biological markers within the tested population (evidence).”

The models used in ACTA are trained with SciBERT, a language model for scientific text, that has been shown to work on texts from different application domains (Beltagy, Lo, & Cohan, 2019). Although the content of articles might differ from clinical trials, the structure of the abstracts is similar, including elements such as background, methods, results, and conclusions. Thus, as arguments can be extracted from abstracts not necessarily dealing with clinical trials and PICO elements detection can be generalized to every biomedical article, we repurposed ACTA to also annotate the articles from the COVID-19 corpus. Thus, we analyzed every abstract and translated the result into RDF to create the COVID-19 Argumentative Knowledge

¹³ <https://github.com/Wimmics/covidontheweb>

¹⁴ PICO is a framework to answer healthcare questions in evidence-based practice that comprises patients/population (P), intervention (I), control/comparison (C), and outcome (O).

Graph (CORD19-AGK), which represent the argumentative components through the Argument Model Ontology (AMO)¹⁵, the SIOC Argumentation Module (SIOCA)¹⁶, and the Argument Interchange Format¹⁷. Further, the PICO elements are described as annotations of the argumentative components in a similar way to the NE and disambiguated against UMLS concepts and semantic types.

3.3. Publishing and Querying the Covid-on-the-Web Data Set

The Covid-on-the-Web data set has a DOI and can be downloaded from Zenodo¹⁸. It can also be queried through our public SPARQL endpoint¹⁹. The RDF data set embeds detailed meta-data describing licensing, authorship, provenance, interlinking, and access information, and the vocabularies used.²⁰ Additional information regarding reproducibility and sustainability have been detailed and discussed in Michel, Gandon et al. (2020).

4. LINKED DATA VISUALIZER

The Linked Data Visualizer is a generic visualization tool for the Semantic Web of Linked Data. It enables the exploration of custom subsets of linked data sets defined via SPARQL queries. Figure 3 provides an overview of the LDViz architecture. It comprises a querying management interface, where users can manage predefined queries, by viewing, editing and visualizing their results, as well as cloning them to create new queries. The interface contains a query editing form, where the user can type their own queries. Upon submitting a query, the obtained results undergo a transformation process, which output data corresponds to the expected format for the visualization. The user can then explore the resulting data using the MGExplorer visualization framework.

In this section, we describe the operational mode of LDViz with particular focus to the querying management and the visualization interfaces. We further demonstrate the versatility of LDViz to explore the Covid-on-the-Web data set through a set of use case scenarios presented in Section 5.

4.1. Query Management Interface

The query management interface (Figure 4) allows users to create and edit their own SPARQL queries. In Figure 4a, we can see the menu that lists and allows managing predefined queries, and Figure 4b–e depict the interface areas enabling the addition and customization of queries. This interface also enables the preview and exporting of a query's results (see Figure 4f). These can be visualized via the MGExplorer graphic library and/or exported as JSON files containing either the results in the SPARQL JSON format or the transformed results used as input to the visualization. The user can type the query in a text area, which can include customizable parameters specified through HTML forms, such as the publication date. Upon submitting a query, the results are processed by a transformation engine that converts the SPARQL JSON format into the JSON format expected by the graphic library.

The transformation engine is generic enough to support the exploration of different variables of the data set. This flexibility allows us to explore graphs with different topologies

¹⁵ <https://purl.org/spar/amo/>

¹⁶ <https://rdfs.org/sioc/argument#>

¹⁷ <https://www.arg.dundee.ac.uk/aif#>

¹⁸ <https://doi.org/10.5281/zenodo.4247134>

¹⁹ <https://covidontheweb.inria.fr/sparql>

²⁰ <https://ns.inria.fr/covid19/covidontheweb-1-2>

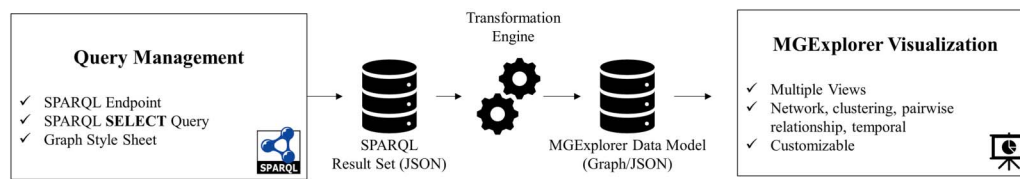


Figure 3. Linked Data Visualizer architecture overview: the Query Management Interface, the Transformation engine, and the Visualization Interface supported by the MGEExplorer visualization tool.

(e.g., with nodes featuring publications, authors, NE). In the context of LDViz, this is made possible by using a SPARQL query that requires at least three variables: $?s$ and $?o$, which describe the nodes (e.g., authors or NE) related by a particular document identified by a variable $?p$. An alternative to $?s$ and $?o$ is the variable $?author$, which contains a list of authors. In addition to these variables, the system allows three other reserved variables that serve to describe the edges ($?p$) of the output graph visualization: $?type$, $?label$, and $?date$. The variable $?type$ can be used to type the edges of the output graph (e.g., by publication type). Due to people's perceptual and cognitive limits regarding visualizations,

a 1. Co-authorship within publications about coronaviruses
Query 2. Looking for papers about cancer and respiratory disease.

b Query Form

Query Title: Co-authorship within publications about coronavirus **c**

SPARQL Endpoint: <https://covidontheweb.inria.fr/sparql>

Custom query variables

Publication Period: From 2015 To 2020 **d**

Named Entities +

Query Prefixes +

- PREFIX pro: <http://purl.org/spar/pro/> **e**
- PREFIX foaf: <http://xmlns.com/foaf/0.1/>
- PREFIX dcterms: <http://purl.org/dc/terms/>
- PREFIX dc: <http://purl.org/dc/elements/1.1/>
- PREFIX schema: <http://schema.org/>

```
SELECT ?p (group_concat(distinct ?name ; separator = '-') as ?author) (sample(?d) as ?date) (sample(?t) as ?label)
WHERE {
  select * where {
    ?p dcterms:title ?t ;
    dcterms:issued ?d ;
    dc:creator ?name .
```

Apply Stylesheet:

f Graph Style sheet

```
{
  "appl": {
    "name": "covid",
    "debug": true
```

g Visualize SPARQL Query Results View SPARQL Query Results View Transformed Results Export SPARQL Query Close Query View

Figure 4. The Query Management Interface. (a) The listing of predefined queries and associated actions. (b) The querying area features: (c) query title and SPARQL endpoint, (d) custom parameters form, and (e) a query editing area. (f) The graph style sheet editing area. (g) The visualization and exporting of results.


```

select ?p (group_concat(distinct ?name ; separator = '--') as ?author)
  (sample(?d) as ?date) (sample(?t) as ?label) where {
  select * where {
    ?doc dct:title ?t ; dct:issued ?d ; dce:creator ?name .
    filter contains(?t, "coronavirus")
    filter (?d >= "$beginYear-01-01"^^xsd:date) # $beginYear = 2015
    filter (?d <= "$endYear-12-31"^^xsd:date) # $endYear = 2021
  } limit 1000
} group by ?p

```

Listing 1. SPARQL query used in Use Case Scenarios 1 and 4 to retrieve the coauthorship network within publications about “coronavirus” between 2015 and 2021.

only a certain number of graphic elements can be drawn on the screen. Thus, we allow the variable `?type` to be bound to only four different values describing the edges. When it is bound to more than four distinct values in the SPARQL query result, the system automatically determines the three most relevant ones based on the number of bindings and classifies the remaining values as “Other.” The `?label` variable allows us to provide a description of the edges in natural language (e.g., the value of `rdfs:label` properties describing resources). Finally, the `?date` variable is used to provide a visual representation of the distribution of edges over time (e.g., publication year).

When dealing with a new data set, researchers often have to debug and test multiple queries to discover the contents of the data set. To ease the customization of queries and the use of the interface by the domain expert, we provide query templates that allow one to interactively define the value of certain parameters, such as the publication period and NE of interest (see Listing 1 for an example).

A Graph Style Sheet language (GSS) serves to transform the default node-link diagrammatic representation through the declarative specification of visibility, layout, and styling rules applied to its nodes and arcs (Pietriga, 2006). Based on this concept, we associate each query to a GSS that the user can edit (see Figure 4e) to customize the resulting node-link diagram (see Listings 2 and 3 for an example). Further to modifying the colors and shape of nodes and edges, we enable, through the GSS, the linking of external services to the visualization interface as a way of extending the analysis. For instance, the Corese engine (Corby, Gaignard et al., 2012) is an RDF processor that enables, among other things, the production of new knowledge through inference rules. Thus, one could include this service on the GSS, which would allow the exploration of the visualized resources through the Corese engine. Further, we can use this feature to support on-the-fly exploration of argumentative graphs of publications identified throughout the visual exploration process by including the ACTA service (see Section 5.5 for more details).

Although we demonstrate the usage of the querying and visualization interfaces for exploring the Covid-on-the-Web dataset, LDViz can be used to query and visualize data from any

```

{"node": { "fst": { "color": "green"}, "snd": { "color": "orange"} },
 "services": [{"label": "ACTA", "url": "http://134.59.134.234:8081/analyseddocs?search="},
 {"label": "Browser Corese", "url": "http://corese.inria.fr/srv/service/covid?uri="}]}

```

Listing 2. Graph Style Sheet used in Use Case Scenarios 2 and 5.

```

# wdt:P279 = subclass of, wdt:P31 = instance of
# wd:Q1134583 = coronavirus family, wd:Q12078 = cancer
prefix wd: <http://www.wikidata.org/entity/>
prefix wdt: <http://www.wikidata.org/prop/direct/>

select distinct ?s ?p ?o ?label ?pmid ?authorList ("fst" as ?style1) ("snd" as ?style2)
from <http://ns.inria.fr/covid19/graph/entityfishing>
from <http://ns.inria.fr/covid19/graph/articles>
from named <http://ns.inria.fr/covid19/graph/wikidata-named-entities-full>
where {
  ?annot1 schema:about ?p ; oa:hasBody ?dis1.
  ?annot2 schema:about ?p ; oa:hasBody ?dis2.
  ?p dct:title ?label ; bibo:pmid ?pmid .
  graph <http://ns.inria.fr/covid19/graph/wikidata-named-entities-full> {
    {?dis1 rdfs:label ?s. filter(?dis1=wd:Q12078)} UNION
    {?dis1 wdt:P279 wd:Q12078; rdfs:label ?s.} UNION {?dis1 wdt:P31 wd:Q12078; rdfs:label ?s.}
    {?dis2 rdfs:label ?o. filter(?dis2=wd:Q1134583)}
    UNION {?dis2 wdt:P279 wd:Q1134583; rdfs:label ?o.} }
  {select ?p (group_concat(?name ; separator="--") as ?authorList) where {
    ?p dce:creator ?name
  } group by ?p}
} limit 1000

```

Listing 3. SPARQL query used in Use Case Scenarios 2 and 5 to retrieve the co-occurrence network within publications of NE related to cancer and coronavirus.

SPARQL endpoint. The querying form contains a field where the user enters the endpoint URL, and the only requirement is that the query returns values for the above-listed predefined set of variables. Hence, what we propose with LDViz is a generic visualization tool for the Semantic Web of Linked Data.


As for any visualization, user queries must be translated to a query language that recovers the necessary data from the database to solve the exploratory task. In this paper, the user queries were identified during interviews with users from INCa and Inserm and translated into SPARQL queries by data scientists. Thus, the query management interface is intended to help expert users (developers and data scientists) to create suitable SPARQL queries for exploring the data set. However, expert users such as biomedical researchers do not need to know SPARQL to visualize and interact with the results of queries. Indeed, they may benefit from a public vitrine²¹ simply by selecting a predefined query to explore the results with MGExplorer without having to deal with SPARQL expressions (Figure 5). The visibility of the predefined queries in the vitrine is settled when queries are created at the query management interface. In the next section, we describe how users can interact with the data resulting from those queries by means of an information visualization interface.

4.2. Visualization Interface

As mentioned earlier, LDViz uses the MGExplorer (**M**ultidimensional **G**raph **E**xplorer) (Menin, Cava et al., 2021) graphic library to support the visual exploration of the Covid-on-the-Web data set. More than a collection of charts, MGExplorer is a visualization tool based on the concept of chained views, which supports the exploration of multidimensional network data,

²¹ Accessible at <https://covid19.i3s.unice.fr:8080/>


Covid Linked Data Visualizer




The goal of this application is to support the analysis and exploration of scientific publications about the Covid-19. The data used in the visualization below come from the endpoint <https://covidontheweb.inria.fr/sparql>

Query Covid Knowledge Graph

Start by selecting a predefined query from the combo box below. To explore the data, click-right over the elements of the graph and select a visualization.

Select a query: Looking for papers about cancer and respi... 

Looking for papers about cancer and respiratory disease.
 Looking for papers about cancer and coronavirus.



[Export SPARQL query](#) [Export SPARQL query result](#) [Save result](#) [About](#)

Figure 5. Public vitrine of Covid-19 Linked Data Visualizer.


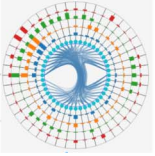
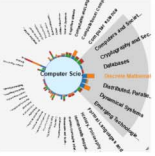

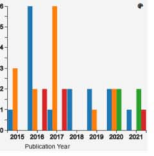

while keeping provenance information to enable further study of users' reasoning based on their interactions with the system. The visual exploration process in MGExplorer consists of two phases, described as follows:

1. the *overview phase*, which consists of visualizing the network defined by the SPARQL query results through a node-link diagram (see description below). This visualization allows the user to get an overall understanding of the clusters within the data; and
2. the *exploratory phase*, where the user can further explore items of interest by selecting them directly on the visualizations, which subsets the data to be explored via a new suitable visualization technique.

The generic aspect of MGExplorer enables the combination of multiple visualizations to support the comparison of two or more different subsets of data through a particular perspective provided by a particular visualization, and the comparison of different perspectives of the same subset of data using multiple, complementary visualization techniques. In particular, we currently support data exploration through six views, summarized in Table 1 and described as follows:

- The **node-link** diagram shows a set of nodes, which represent data items (e.g., authors), and their relationships represented through line segments connecting them. In MGExplorer, this visualization technique provides an overview of the relationships within items of the input data. In our use case scenarios (Section 5), the relationships are defined by scientific publications, either to reveal coauthorship networks or the co-occurrence of NE.
- The **ClusterVis** technique (Cava, Freitas et al., 2017) enables the inspection of clusters and data attributes (e.g., publication type) within the subset of items (e.g., authors or NE). The visualization has a multiring layout, where the innermost ring is formed by dots representing data items, and the remaining rings display the data attributes, which can be customized and reordered by the user. The items in the innermost ring that

Table 1. Classification of visualization techniques available in MGExplorer according to the type of analysis they provide

Node-link diagram  network	ClusterVis  clusters	IRIS  pairwise	GlyphMatrix  pairwise	Bar chart  distribution	Listing  listing
---	---	--	--	--	---

belong to the same subcluster are connected via curved lines, which one can highlight by hovering over the items. The remaining rings are formed by bars where height and color encode different data attributes (e.g., the height encodes count and the color encodes the types of publications of a specific author).

- The **IRIS** technique represents the pairwise relationships between an item of interest (e.g., an author) and the remaining items in a particular subset of data, which relationship is described by data attributes (e.g., publication count and type) (Cava, Freitas et al., 2014). This technique is inspired by the eye’s iris, which can only focus on a certain amount of information at the time (i.e., what is visible within our field of view). The selected item is represented in the IRIS as a circle at the center of the view, surrounded by its related items, which are displayed in a way that the ones in the field of view (gray area) are larger than the ones outside this zone, easing information extraction. The user can place any item in the field of view by clicking on it, switching the focus of the IRIS. To represent data attributes describing those pairwise relationships, we use the height and color of a bar placed in between the item of interest and each of its related items.
- The **GlyphMatrix** technique (Cava & Freitas, 2013) features a matrix where rows and columns represent data items (e.g., authors or NE), and the intersection cell between each pair of items contains a glyph encoding the data attributes describing that relationship. The default glyph is based on a radar chart, where each axis displays the count of a different data attribute (e.g., publication type). The technique supports sorting of rows and columns to facilitate information extraction, and hovering over cells to make the glyph larger and more visible through a tooltip feature. This visualization technique could be seen as a combination of the ClusterVis and IRIS by displaying the relationship between an item of interest and other items in a pairwise manner, as well as the relationships within the remaining items in the group.
- The **Bar chart** technique shows the distribution of publications according to a given variable. In our case study, the x-axis encodes temporal information, and the y-axis encodes the counting of publications. The data are displayed as a single bar per time-period or multiple colored bars to represent categorical information of attributes.
- The **Listing** technique lists the items that form the relationship between two or more nodes in the graph. In our case study, it displays the list of publications coauthored by two or more authors or the publications where two or more NE co-occur, according to the subset of data being explored. Each item of the list contains a link to a descriptive web page of the publication, where the user can obtain more information about it. Furthermore, if enabled by the GSS, each item contains a context menu to enable further exploration using an external service (e.g., ACTA).

Downloaded from http://direct.mit.edu/gss/article-pdf/2/4/1301/2007990/gss_a_00164.pdf by guest on 07 September 2023

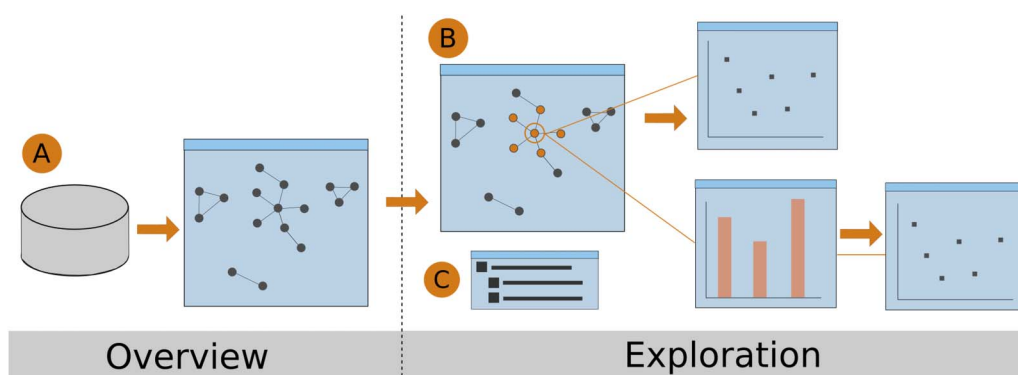


Figure 6. Overview of MGExplorer. Panel A: The node-link diagram provides an overview of the data set. Panel B: Filtering operations enable further exploration of items/subsets of interest through different visualization techniques. Panel C: A history panel records users' actions throughout the exploration process. Image retrieved from Menin, Cava et al. (2021).

Each view is a self-contained element, which includes a visualization technique and supports subsetting operations, enabling further exploration of subsets of data through different views. The views can be dragged, allowing the user to rearrange the visualization space in meaningful ways to the ongoing analysis. They are connected via line segments, which reveal their dependencies and enable tracing back the exploration path, thus preserving provenance information.

Upon submitting a SPARQL query in the query management interface, the data goes through a transformation process, and MGExplorer self-starts with the overview phase. The node-link diagram and a History panel (Figure 6C) are visible during the whole exploration. The history panel displays the exploration path in a hierarchical format to indicate the dependencies between views, and supports quick recovery of the multiple analytical paths that emerge from a particular view. The history panel allows the user to clean the visualization space while focusing on what is relevant to the ongoing analysis by hiding currently displayed visualizations and/or showing any of the previous visualizations.

5. USE CASE SCENARIOS

In this section we illustrate the usage of COVID LDViz to explore the Covid-on-the-Web data set. The goal is to demonstrate what kind of data one can explore using this interface and how the data processing between the query management and the visualization interfaces support a multiperspective exploration of the dataset.

5.1. Scenario 1: Clusters Visualization

Based on the premise that COVID-19 has increased the collaboration between researchers from diverse disciplines around the world (Naujokaitytė, 2021), a biomedical researcher from INCa was interested on searching for information about existing collaborations on the theme of the relationship between COVID-19 and cancer (or more generally between COVID-19 and other diseases) in order to analyze the nature of these collaborations, their impact, and their evolution. In this scenario, we illustrate how LDViz could assist this analysis by exploring coauthorship networks.

We use a subset of data describing the coauthorship network within publications related to coronavirus families retrieved with the query presented in Listing 1, which resulted in 4,238 RDF triples corresponding to publications having the word “coronavirus” in the title. These

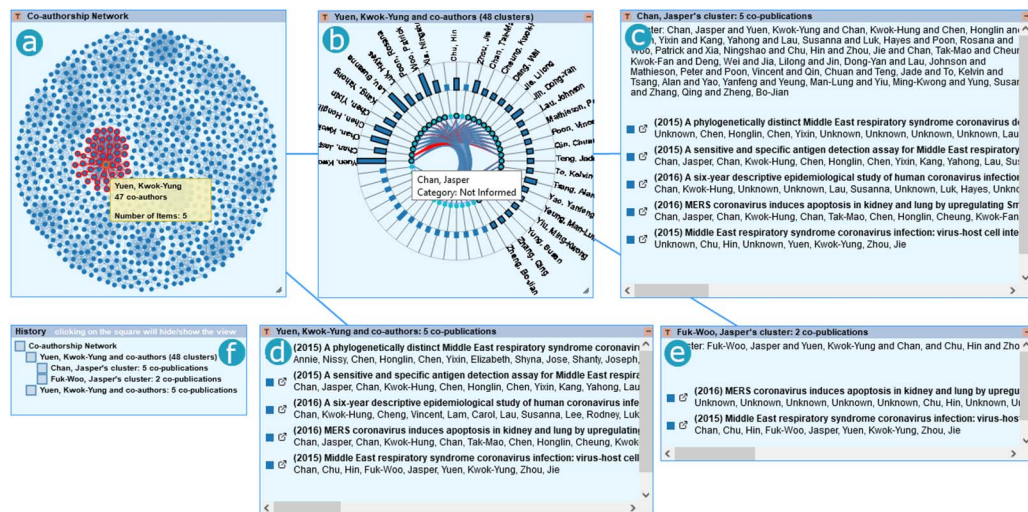


Figure 7. Exploratory path of Scenario 1. (a) We use the NodeEdge diagram to identify an author of interest for exploration. (b) The ClusterVis reveals the subclusters within the set of coauthors and their copublications. (c)–(e) The views depict the publications produced within each subcluster. (e) The total publications of the author of interest. (f) The history shows which charts were opened, their order, and inner dependencies.

results were then transformed into a graph with 879 nodes (authors) and 4,053 edges (connections between authors). Figure 7 depicts the exploratory path that we follow during this scenario, which illustrates how one can explore clusters of coauthors and related information to their copublications. As mentioned earlier, the MGExplorer visualization interface self-starts with an overview of coauthorship clusters through the node-link diagram and the history tree of the exploratory process, which is progressively completed based on the user's interactions.

In the node-link diagram, we identify a dense subgraph related to the author Yuen, Kwok-Yung (Figure 7a), who will be our author of interest for this exploration. We hover over the node representing the author, where we observe that they have 47 coauthors, with whom five scholarly articles have been published. Subsequently, we right-click on the node to activate a context menu that allows subsetting the data and explore it with another visualization technique. We choose the ClusterVis view, where we can explore the different clusters within the subset of coauthors selected in the node-link (Figure 7c). For two different clusters, we subset the data by hovering over a particular author and display the list of publications which they coauthored together (Figure 7d–e). Finally, we could compare the contributions made within those clusters and the complete list of publications coauthored by our author of interest (Figure 7f), to understand the impact of these coauthorship relationships in terms of the number and quality of publications they have together.

5.2. Scenario 2: Customizing the Graph Topology

The generic structure of LDViz allows the construction of graphs with different topologies. The user can choose the variables that correspond to the nodes and the connection between them (e.g., in the previous scenario, nodes correspond to a variable that describes the authors' names and the edges correspond to a variable that describe the documents they coauthored). Together with biomedical researchers, we have identified the task “to identify the articles that mention both a type of cancer and a virus of the corona family” as being relevant for their analyses. Thus, in this scenario, we illustrate how we can use LDViz to solve this domain-related task. Using the query presented in Listing 3, we retrieve the RDF triples that correspond

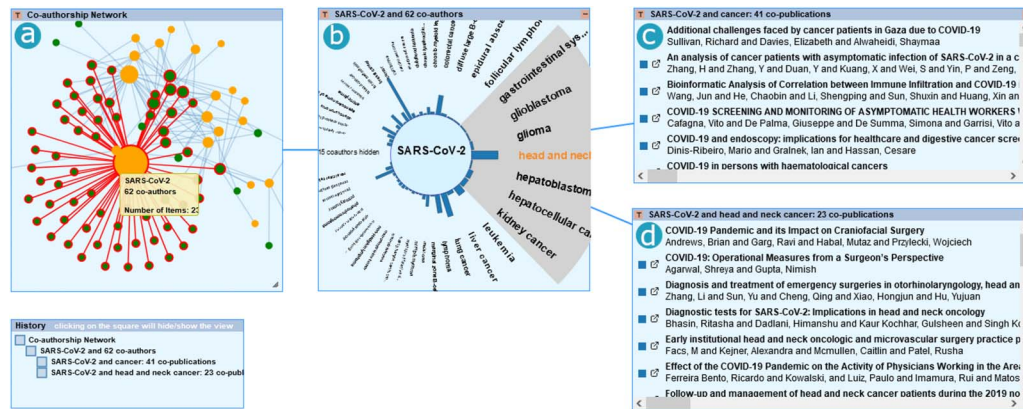


Figure 8. Exploratory path of Scenario 2. (a) In the node-link diagram we see the connection between types of cancer (green) and viruses from the coronavirus family (orange). (b) The IRIS shows relationship between SARS-CoV-2 and different types of cancer in a pairwise manner. (c) The list of publications related to SARS-CoV-2 and cancer in general, and (d) head and neck cancer.

to the pattern $?s \rightarrow ?p \rightarrow ?o$, where $?s$ and $?o$ are, respectively, NE related to (i.e., equal to, subclass of, or instance of) “cancer” and “coronavirus” NE, and $?p$ refers to the publications that contain these NE on their text body. The relationships are determined by publications; however, unlike the Scenario 1, this query modifies the topology of the graph to represent the relationships between NE instead of coauthors.

Figure 8 depicts the exploratory path followed in this scenario to solve the above-described domain-related task. We explore a data set that contains 452 RDF triples, which results in a graph with 94 nodes and 169 edges. Because in this data set, we deal with two types of nodes (i.e., related to either “cancer” or “coronavirus”), we use the GSS feature (see Listing 2) to color these different types of nodes accordingly (i.e., green encodes cancer and orange encodes coronavirus), easing the visual identification of the relationship between the cancer- and coronavirus-related nodes directly in the node-link diagram (Figure 8a). Due to the nature of the data, we can easily spot a large subgraph originating from the SARS-CoV-2 named entity, which is associated with 62 types of cancer through 232 publications. We further explore the subset of data within this subgraph by right-clicking on the node representing SARS-CoV-2 and choosing the IRIS visualization, which displays the relationships of this named entity with the different types of cancer in a pairwise manner (Figure 8b). We could observe via the longest bar in the IRIS that SARS-CoV-2 mostly co-occurs with “cancer” in 41 publications (Figure 8c); which types are not specified. Further, we observe that the second most recurrent co-occurrence of SARS-CoV-2 is with “head and neck cancer,” for which we observe the existence of 23 publications (Figure 8d). The Listing view displays the publications together with links to their descriptive pages in the Covid-on-the-Web data set, where the user can find more information about each document²².

5.3. Scenario 3: Exploring Data Attributes

The previous exploration scenarios allow the user to see the relationship between coauthors or NE, which can be characterized by the number of related publications. Thus, this scenario illustrates how we can use LDViz to explore custom data attributes of a coauthorship network within coronavirus-related publications. In particular, we will use a data set that describes

²² Example of a document descriptive page in the Covid-on-the-Web dataset: <https://covidontheweb.inria.fr/describe?url=http://ns.inria.fr/covid19/28ecacb70247f4fb6a4923a99d0905153c23f88a>

```

select ?p (group_concat(distinct ?name ; separator = '--') as ?author)
  (sample(?d) as ?date) (sample(?t) as ?label)
  (sample(?label) as ?type) where {
select * where {
  ?p dct:title ?t ; dct:issued ?d ; dce:creator ?name .
  filter contains(?t, "coronavirus")

  graph <http://ns.inria.fr/covid19/graph/entityfishing> {
    ?a1 a oa:Annotation; schema:about ?p ; oa:hasBody ?uri .
    ?uri rdfs:label ?subject .
    FILTER (langMatches( lang(?subject), "EN" ) )
  }
} limit 10000
} group by ?p

```

Listing 4. SPARQL query used in Use Case Scenario 3 to retrieve the coauthorship network within publications about “coronavirus” described by research subject.

publications through the research topic retrieved with the query presented in Listing 4. In the context of the Covid-on-the-Web dataset, this information originates from the `schema:about` property, which refers to a set of NE that can be used to describe the research topic of the publication. The resulting data set has 1,265 RDF triples, which were transformed in a graph with 356 nodes (authors) and 1,262 edges (copublications). From the resulting data, the system identified the values “sequence alignment,” “reverse transcriptase,” and “transfection” as the most relevant research topics to describe the publications within the data and classified the remaining under the “other” category.

Figure 9 depicts the exploratory path of this scenario. We inspect the clusters of coauthorship within the associations of different authors through the ClusterViz visualization. We can observe, for instance, that the researcher Collisson, Ellen (Figure 9a) has publications about different topics (i.e., sequence alignment and other) within different clusters of coauthorship, and the publications coauthored by Chu, Chen-Chung (Figure 9b) refer to the “other” category of topics and are distributed throughout different clusters of coauthorship. Finally, we

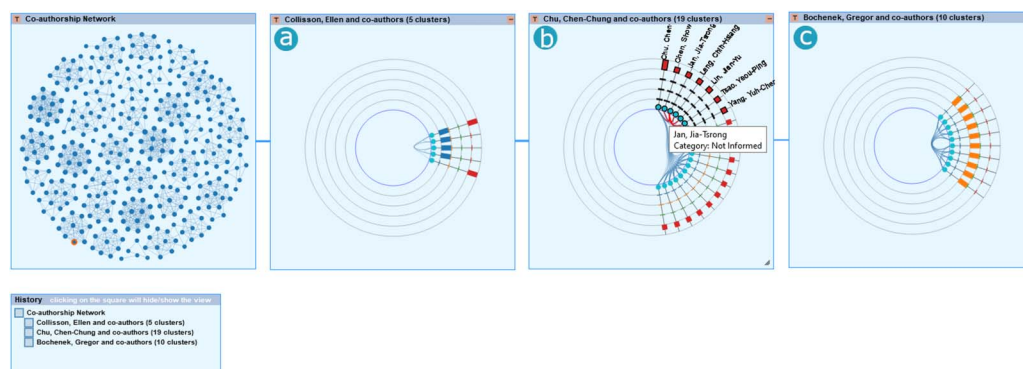


Figure 9. Exploratory path of Scenario 3. (a)–(c) The ClusterViz visualizations depicts the clusters of different authors, where we see their collaborations in different research topics (blue encodes “sequence alignment,” green encodes “reverse transcriptase,” and orange encodes other subjects).

observe that the publication coauthored by Bochenek, Gregor (Figure 9c), for instance, refers to the topic of “reverse transcriptase.”

5.4. Scenario 4: Exploring the Temporal Aspect of Relationships

Studying the evolution over time of coauthor relationships or NE co-occurrence could help understand when collaborations between authors were stronger or when certain research topics were of greater interest, which information could be further explained with context (e.g., nowadays the research around the coronavirus topic is stronger than ever due to the COVID-19 pandemic). Thus, in this scenario, we illustrate how one can use the LDViz interface to explore the temporal aspects of relationships, particularly coauthorship within publications related to coronaviruses (see Listing 1).

Figure 10 depicts the exploratory path used in this scenario. Similar to Scenario 1, we use the node-link diagram to identify the author with the most coauthors (i.e., Yuen, Kwok-Yung; hereafter called *author A*) with 47 coauthors associated through five publications (Figure 10a). We further explore the relationship between author A and their coauthors through the GlyphMatrix visualization, which shows the types and number of copublications between author A and every other coauthor, as well as the copublications among author A’s coauthors. By ordering rows and columns by the number of copublications, we can observe in the GlyphMatrix that author A’s most recurrent coauthor is Lau, Susanna (hereafter called *author B*) (Figure 10b), with whom they have four publications. Thus, to verify when these collaborations happened, we explore the temporal distribution of copublications between those authors by subsetting the data in the GlyphMatrix visualization and exploring it on the Histogram technique (Figure 10c). We observe that they had collaborations in 2015 and

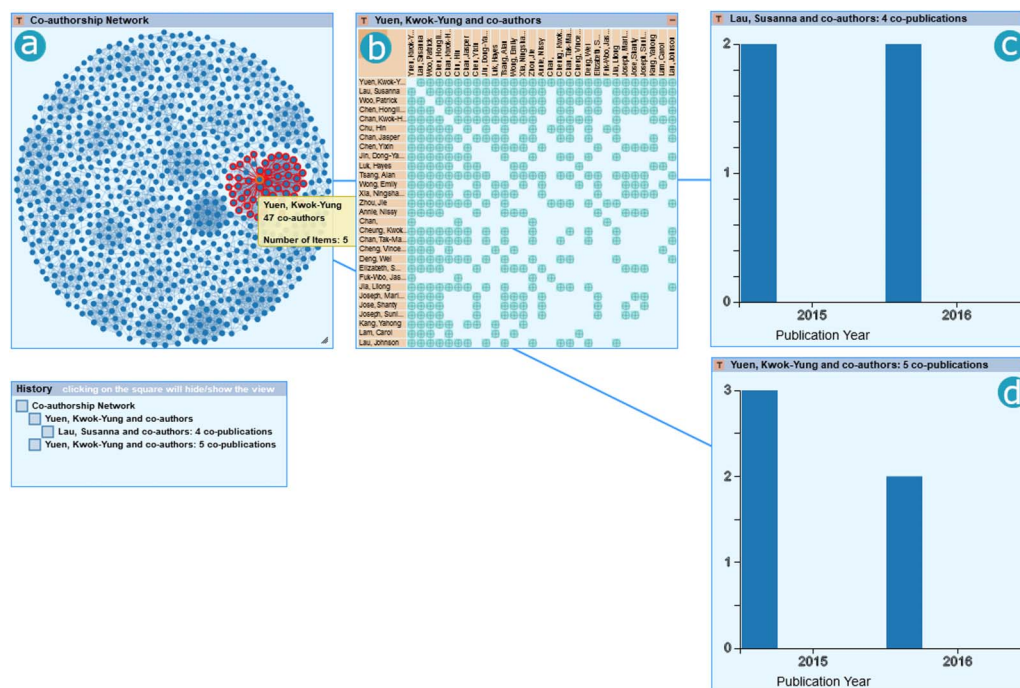


Figure 10. Exploratory path of Scenario 4. (a) We identify on the NodeEdge diagram the author of interest. (b) In the GlyphMatrix, we identify their most recurrent coauthor at the top-left cells, and we (c) explore the temporal distribution of their copublications using the Histogram, which we compare with (d) the temporal distribution of publications coauthored by the author of interest.

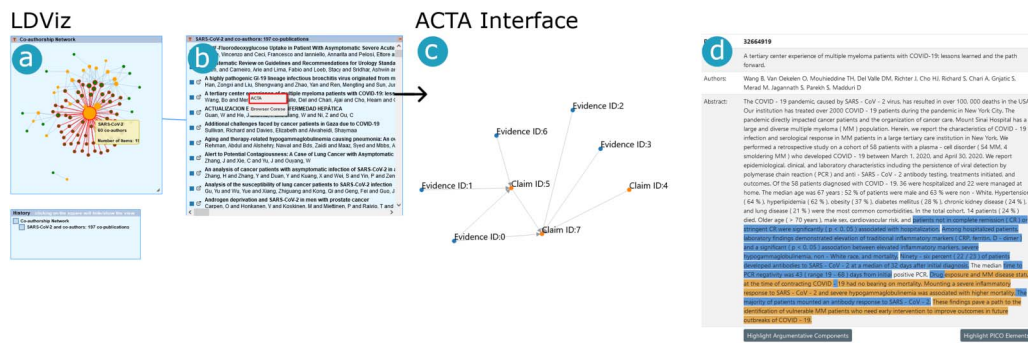


Figure 11. The exploratory path of Scenario 5. In the LDViz interface we (a) find a node of interest, and (b) explore its related publications through the Papers List view. We right-click on a document and explore it using the ACTA interface, where we can (c) visualize the argumentative graph and (d) explore where the claims, evidence and PICO elements appear in the document's abstract.

2016. When comparing to the totality of copublications related to author A (Figure 10d), we observe that four out of five publications are coauthored by author B, which could indicate a strong collaboration between those authors in copublications related to the coronavirus topic. We can also observe that this collaboration appears to have ended 5 years ago, as the data set contains publications from 2015 to 2021.

5.5. Scenario 5: Exploring Argumentation Graphs with the ACTA Interface

As mentioned earlier, the GSS feature allows the user to include external services in LDViz, such as a service that enables further exploration of the resources currently being visualized with the LDViz interface. In this scenario, we explore the subset of data used in Scenario 2 (i.e., the set of publications where NE related to “cancer” and “coronavirus” co-occur) to illustrate how one can use the ACTA interface to visualize the argumentative graph of a certain publication identified during the exploratory process. As one can see in Listing 2, the GSS form associated with the query contains an object called “services” that provides the redirection information for the ACTA interface (i.e., a call to “https://134.59.134.234:8081/analyseddocs?search=”). The documents used in the Covid-on-the-Web data set often originate from the PubMed archive²³, where each document has a unique identifier. Thus, upon the selection of a document, the LDViz system launches the ACTA service by redirecting the user to the given URL, while providing the document identifier as a parameter.

Figure 11 depicts the exploratory path used in Scenario 5. As for Scenario 2, we identify the larger subgraph in the node-link diagram, which is the one connecting to the node that corresponds to the named entity “SARS-Cov-2” (Figure 11a). Using the Histogram, we display the 232 publications where this named entity occurs (Figure 11b). Subsequently, we can choose any of the listed publications for which we would like to visualize the argumentative graph using ACTA. We right-click on the publication of interest and choose the “ACTA” option on the context menu that appears. This action redirects the user to the ACTA interface, which retrieves the selected document from the PubMed server, analyzes it, and displays the resulting argumentative graph with the relationships between claims and evidence, and PICO elements (Figure 11c). One can also inspect these elements using the textual information (Figure 11d), where we can choose to highlight the argumentative sentences or the PICO elements. Alternatively, one can query the CORD19-AGK²⁴ data set to explore claims and evidence graph

²³ <https://pubmed.ncbi.nlm.nih.gov/>

²⁴ <https://ns.inria.fr/covid19/graph/acta>

related to one or more publications directly on LDViz by using a SPARQL query where $?s$ and $?o$ correspond to claims and evidence, and the $?p$ variable corresponds to the publication(s) where they were identified.

6. DISCUSSION

The Covid-on-the-Web project integrates knowledge from diverse research areas (i.e., semantic web, NLP, and visualization) to assist researchers, particularly in the biomedical field, to explore the COVID-19 scientific literature. For this purpose, we created a linked data version of the COVID-19 data set and enriched it via entity linking and argument mining. To the best of our knowledge, the Covid-on-the-Web data set is the first public knowledge graph on the Web integrating publication metadata, NE, arguments, and PICO elements into a single, coherent whole. The openness aspect of our data set and code should enable contributors to advance the current state of knowledge on this disease. Further, we believe the Covid-on-the-Web data set could serve as a foundation for Semantic Web applications and benchmarking algorithms.

Moreover, we proposed a set of visualization interfaces to assist in the exploration of the Covid-on-the-Web data set from different perspectives, enabling the resolution of various domain-related questions. In this paper, we have particularly focused on the LDViz visualization tool, which supports the visual exploration of subsets of data defined by SPARQL queries. The tool is based on the MGExplorer visualization framework, which proposes a collection of charts linked together through a chained visualization approach that allows us to keep track of the exploration path, assisting with the understanding of the sense-making process. This visualization aims to help users understand the relationships within the results: For example, users can run a query to visualize a coauthorship network; then use IRIS and ClusterVis to understand who is working together and on which research topics. An interesting aspect of our approach is that one can change the graph topology to explore relationships between different kinds of items. For instance, the user could execute a query that looks for papers mentioning the COVID-19 and diverse types of cancer, as illustrated in Use Case Scenario 2 (see Section 5.2). Another strong aspect of LDViz relies on the possibility of exploring the relationships within any subset of data originating from any SPARQL endpoint thanks to the data transformation engine that adapts the query's results to the data format required by the visualization.

In addition to our partners from the Inserm and INCa institutes, the resources and services proposed in the Covid-on-the-Web project have aroused the interest of other institutions, such as Antibes and Nice Hospital. In particular, we have shown in this paper that our approach supports the different types of analyses evoked by domain users: the analysis of clinical trials to make evidence-based decisions, which we support via argumentative graphs; the study of the relationship between coronaviruses and other diseases, such as cancer, which we provide through co-occurrence networks that assist their search for scientific articles on the topic; and the identification of researchers, institutions, or countries working on the topic via coauthorship network analysis.

Although a first level of evaluation is shown by translating the user queries to SPARQL queries to visual data in LDViz, which shows that our data set and visualization services support the resolution of users' queries, user evaluations are essential to validate the usability and utility of a visualization. However, evaluating LDViz (as well as any visualization) is not a trivial task because it has been designed to support exploratory tasks, which are the hardest ones to replicate in an experiment (Ellis & Dix, 2006). Furthermore, the value of LDViz can only be assessed when used by professionals on the application domain (e.g., biomedical researchers),

who are difficult to recruit because they are not necessarily available to take part in experiments. Future work includes implementing user-based evaluations to investigate the usability of LDViz tool for exploring linked data sets in general, and in particular its suitability for analyzing the COVID-19 scientific literature and assisting in the resolution of domain-related tasks.

The generic aspects of our tools allow us to later on apply the resources to a wider set of use case scenarios, which possibility has been evoked by our biomedical partners, who would like to perform similar analyses over issues other than the COVID-19. In fact, the LDViz interface has been applied to two other publication data sets (i.e., HAL open archive²⁵ and the Microsoft Academic Knowledge Graph²⁶, for which a set of predefined queries are available at <https://covid19.i3s.unice.fr:8080/hal>). The genericity of our approach enables the exploration of data from any SPARQL endpoint, such as DBpedia²⁷, from which we explored the ontology and RDF Schema information, as well as a costarring relationship using movie information²⁸. The tool also has a generic service that enables the querying and visualization of any SPARQL endpoint, which URL can embed a SPARQL query and the URL of a SPARQL endpoint²⁹, to directly visualize the resulting data. Furthermore, from a linked data perspective, one can use the Corese SPARQL service³⁰ to combine data from different SPARQL endpoints using federated queries.

Typically, in an exploratory visualization, the user has no defined goal and is looking for no particular outcome (Leng, 2011). Although, in context of LDViz, the user does have an initial query and would, therefore, have an exploratory goal in mind, throughout the exploratory process one can make new discoveries that might not be directly related to the initial query but that could be equally interesting. The user could yet be interested in exploring the same data through different visualization techniques, which could provide them with a different perspective on the data and would create an alternative exploratory path to solve the same query. In this context, because visualization can help to recall, revisit, and reproduce the sense-making process through visual representations of provenance data, MGExplorer visually represents the dependencies between views through line segments and uses the history panel to display exploratory actions hierarchically, retaining parenting and visualization information such as the data and technique used. The interactive aspect of the history panel allows the user to trace back their exploratory path, while allowing them to start an alternative exploratory path from a given point in history. Future work includes implementing querying support for alternative data sets through a mechanism of follow-up queries, which allows users to launch a new query based on an item or subset of items of interest identified in a view, bringing together complementary data from external data sets to enrich the analysis.

A strong aspect of the LDViz interface, and in particular, the MGExplorer visualization tool, is the ability to record and visualize provenance information. Currently, this information is restricted to the subsets of data and the visualizations used during the analysis. Thus, we also intend to increase the variety of provenance information we record, considering the several interactions used during the exploration (e.g., clicks, hovering, data sorting, etc) that might be relevant to understanding users' reasoning, as well as to include a feature that allows users to

²⁵ <https://data.archives-ouvertes.fr/doc/sparql>

²⁶ <https://makg.org/sparql>

²⁷ <https://fr.dbpedia.org/sparql>

²⁸ Available at <https://covid19.i3s.unice.fr:8080/ldviz>

²⁹ <https://covid19.i3s.unice.fr:8080/ldviz?query=<SPARQL query>&url=<SPARQL endpoint URL>>

³⁰ <https://corese.inria.fr/sparql>

make annotations throughout the process regarding the historic items. Future work also includes the analysis of the resulting provenance data. For instance, we could analyze the resulting data to identify the most common usages of the system (standard choices of visualizations and instantiating order) according to different types of tasks, which could be used to introduce the system to new users, suggest some well-known workflows of analysis, and to improve overall user experience. Furthermore, we could validate these usage patterns through user-based evaluations involving experts in the application domain, who would evaluate whether and at which level the common detected workflows respond to their needs and how it could be improved (i.e., which alternative exploratory path one would follow to solve specific user queries).

For the purpose of extending the range of resources and services of the Covid-on-the-Web project and, thus, extending and improving the supported types of analyses, future work includes integrating new visualization services, such as ARViz (Menin, Cadorel et al., 2021), which allows the visual exploration of association rules describing patterns of co-occurring NE within publications through three complementary visualization techniques: a scatter plot, a chord diagram, and an association graph³¹. The tool currently works separately with a pretreated subset of data extracted from the Covid-on-the-Web data set. However, the association mining algorithm can process any RDF data set, the results of which could then be explored with ARViz. Thus, future work includes the integration of this visualization interface in the LDViz tool, where the user could analyze and explore meaningful data defined via SPARQL queries, similarly to what is done with the MGExplorer, resulting in a completely integrated tool for extracting and exploring knowledge from scientific literature through various perspectives.

7. CONCLUSION

In this paper, we presented the data set and software resources provided by the Covid-on-the-Web project, with a particular focus on the visualization services proposed to support the exploration of the COVID-19 scientific literature. Based on the needs of biomedical researchers, who are partners of the project, we designed and published a linked data knowledge graph describing the NE mentioned in the articles of the COVID-19 corpus and the argumentative graphs they include. The knowledge graph generation pipeline has been published to allow the scientific community to reuse, enrich, and adapt both the data set and the pipeline in meaningful ways to assist users' needs.

Furthermore, we described and demonstrated the use of LDViz, a visualization interface dedicated to the exploration of linked data, which is based on a SPARQL querying interface and the MGExplorer interface, a generic visualization framework designed to explore multidimensional graph data. We have shown the potential of this interface to explore different perspectives to the Covid-on-the-Web data set, supporting the resolution of diverse domain-related tasks.

Future work includes evaluating our resources and services with the participation of expert users in the biomedical domain in terms of usability and suitability to solve the domain-related tasks; developing a querying feature that allows us to dynamically import data into the exploratory process from external data sets, aiming to enrich the ongoing analysis and explore on-the-fly hypotheses; studying provenance information aiming to improve user experience and

³¹ Available at <https://covid19.i3s.unice.fr:8080/arviz/>

the visualization's effectiveness; and integrating new visualization services to extend the support for different domain-related tasks.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of Valentin Ah-Kane and Mathieu Simon and our research partners of Inserm and INCa institutions. We also acknowledge the contribution of Carla Freitas and Ricardo Cava for the initial work on the MGExplorer framework.

AUTHOR CONTRIBUTIONS

Aline Menin: Conceptualization, Investigation, Methodology, Software, Writing—original draft, Writing—review & editing. Franck Michel: Data curation, Investigation, Resources, Software, Writing—review & editing. Fabien Gandon: Funding acquisition, Writing—review & editing. Raphaël Gazzotti: Resources, Writing—review & editing. Elena Cabrio: Supervision. Olivier Corby: Software. Alain Giboin: Investigation, Methodology, Writing—review & editing. Santiago Marro: Resources. Tobias Mayer: Resources. Serena Villata: Supervision. Marco Winckler: Conceptualization, Formal analysis, Methodology, Supervision, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

This work is partly funded by the French government labeled PIA program under its IDEX UCAJEDI project (ANR-15-IDEX-0001) and the 3IA Côte d'Azur (19-P3IA-0002) as well as the CovidOnTheWeb project funded by Inria.

REFERENCES

- Ambavi, H., Vaishnav, K., Vyas, U., Tiwari, A., & Singh, M. (2020). CovidExplorer: A multi-faceted AI-based search and visualization engine for COVID-19 information. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 3365–3368). <https://doi.org/10.1145/3340531.3417428>
- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Pretrained language model for scientific text. *EMNLP, arXiv preprint arXiv:1903.10676*. <https://doi.org/10.18653/v1/D19-1371>
- Bras, P. L., Gharavi, A., Robb, D. A., Vidal, A. F., Padilla, S., & Chantler, M. J. (2020). Visualising COVID-19 research. *arXiv preprint arXiv:2005.06380*.
- Cava, R., & Freitas, C. D. S. (2013). Glyphs in matrix representation of graphs for displaying soccer games results. *The 1st Workshop on Sports Data Visualization. IEEE, 13, 15*. <https://workshop.sportvis.com/papers/cavaSoccerMatches.pdf>
- Cava, R., Freitas, C. M. D. S., & Winckler, M. (2017). ClusterVis: Visualizing nodes attributes in multivariate graphs. *Proceedings of the Symposium on Applied Computing* (pp. 174–179). <https://doi.org/10.1145/3019612.3019684>
- Cava, R., Freitas, C. M., Barboni, E., Palanque, P., & Winckler, M. (2014). Inside-in search: An alternative for performing ancillary search tasks on the web. *2014 9th Latin American Web Congress* (pp. 91–99). <https://doi.org/10.1109/LAWeb.2014.21>
- Corby, O., Gaignard, A., Faron-Zucker, C., & Montagnat, J. (2012). KGRAM versatile data graphs querying and inference engine. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. <https://dl.acm.org/doi/10.5555/2457524.2457672>
- Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. *Proceedings of the 9th International Conference on Semantic Systems* (pp. 121–124). <https://doi.org/10.1145/2506182.2506198>
- Ellis, G., & Dix, A. (2006). An explorative analysis of user evaluation studies in information visualisation. *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (pp. 1–7). <https://doi.org/10.1145/1168149.1168152>
- Fonseca, B. d. P. F. e., Sampaio, R. B., de Araújo Fonseca, M. V., & Zicker, F. (2016). Co-authorship network analysis in health research: Method and potential use. *Health Research Policy and Systems, 14*(1), 1–10. <https://doi.org/10.1186/s12961-016-0104-5>, PubMed: 27138279
- Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., ... West, J. (2020). SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668*. <https://doi.org/10.1101/2020.05.23.112284>

- Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N. T., Yao, Y., ... Szekely, P. (2020). KGTK: A toolkit for large knowledge graph manipulation and analysis. *The Semantic Web – ISWC 2020* (pp. 278–293). https://doi.org/10.1007/978-3-030-62466-8_18
- Jonquet, C., Shah, N. H., & Musen, M. A. (2009). The open biomedical annotator. *Summit on Translational Bioinformatics, 2009* (p. 56). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041576/>. PubMed: 21347171
- Leng, J. (2011). *Handbook of research on computational science and engineering: Theory and practice* (Vol. 2). IGI Global. <https://doi.org/10.4018/978-1-61350-116-0>
- Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2016). Visualizing ontologies with VOWL. *Semantic Web, 7*(4), 399–419. <https://doi.org/10.3233/SW-150200>
- Mayer, T., Cabrio, E., & Villata, S. (2019). ACTA a tool for argumentative clinical trial analysis. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 6551–6553). <https://doi.org/10.24963/ijcai.2019/953>
- Menin, A., Cadorel, L., Tettamanzi, A., Giboin, A., Gandon, F., & Winckler, M. (2021). ARViz: Interactive visualization of association rules for RDF data exploration. *25th International Conference Information Visualisation*. <https://doi.org/10.1109/IV53921.2021.00013>
- Menin, A., Cava, R., Freitas, C. M. D. S., Corby, O., & Winckler, M. (2021). Towards a visual approach for representing analytical provenance in exploration processes. *25th International Conference Information Visualisation*. <https://doi.org/10.1109/IV53921.2021.00014>
- Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., ... Winckler, M. (2020). Covid-on-the-Web: Knowledge graph and services to advance COVID-19 research. In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (pp. 294–310). Springer. https://doi.org/10.1007/978-3-030-62466-8_19
- Naujokaitytė, G. (2021). COVID-19 triggered unprecedented collaboration in research. <https://sciencebusiness.net/covid-19/news/covid-19-triggered-unprecedented-collaboration-research> (accessed July 6, 2021).
- Oniani, D., Jiang, G., Liu, H., & Shen, F. (2020). Constructing co-occurrence network embeddings to assist association extraction for COVID-19 and other coronavirus infectious diseases. *Journal of the American Medical Informatics Association, 27*(8), 1259–1267. <https://doi.org/10.1093/jamia/ocaa117>, PubMed: 32458963
- Pietriga, E. (2006). Semantic web data visualization with graph style sheets. *Proceedings of the 2006 ACM Symposium on Software Visualization* (pp. 177–178). <https://doi.org/10.1145/1148493.1148532>
- Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., ... Mungall, C. J. (2021). KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response. *Patterns, 2*(1), 100155. <https://doi.org/10.1016/j.patter.2020.100155>, PubMed: 33196056
- Sukla, A., Naskar, A., Goel, T., Sangwan, S., Rai, A., ... Dey, L. (2021). Concept driven search and visualization system for exploring scientific repositories. *8th ACM IKDD CODS and 26th COMAD* (pp. 395–399). <https://doi.org/10.1145/3430984.3430991>
- Tu, J., Verhagen, M., Cochran, B., & Pustejovsky, J. (2020). Exploration and discovery of the COVID-19 literature through semantic visualization. *arXiv preprint arXiv:2007.01800*. <https://doi.org/10.18653/v1/2021.naacl-srw.11>
- Verspoor, K., Šuster, S., Otmakhova, Y., Mendis, S., Zhai, Z., ... Martinez, D. (2020). COVID-SEE: Scientific Evidence Explorer for COVID-19 related research. *arXiv preprint arXiv:2008.07880*. https://doi.org/10.1007/978-3-030-72240-1_65
- Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R. M., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., ... Kohlmeier, S. (2020). COVID-19: The Covid-19 Open Research Dataset. *ArXiv, abs/2004.10706*.