Covid-on-the-Web: Exploring the COVID-19 Scientific Literature through Visualization of Linked Data from Entity and Argument Mining

Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, et al.

Covid-on-the-Web: Exploring the COVID-19 Scientific Literature through Visualization of Linked Data from Entity and Argument Mining

Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, Olivier Corby, Alain Giboin, Santiago Marro, Tobias Mayer, Serena Villata, and Marco Winckler

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

Author Note

Aline Menin https://orcid.org/0000-0002-9345-3994

Franck Michel https://orcid.org/0000-0001-9064-0463

Fabien Gandon https://orcid.org/0000-0003-0543-1232

Raphaël Gazzotti https://orcid.org/0000-0002-5618-9776

Elena Cabrio https://orcid.org/0000-0001-9374-7872

Olivier Corby https://orcid.org/0000-0001-6610-0969

Alain Giboin https://orcid.org/0000-0003-1007-0101

Santiago Marro https://orcid.org/0000-0001-6220-0559

Tobias Mayer https://orcid.org/0000-0002-4935-4710

Serena Villata https://orcid.org/0000-0003-3495-493X

Marco Winckler https://orcid.org/0000-0002-0756-6934

Corresponding author: Aline Menin (aline.menin@inria.fr)

Abstract

The unprecedented mobilization of scientists, consequent of the COVID-19 pandemics, has generated an enormous number of scholarly articles that is impossible for a human being to keep track and explore without appropriate tool support. In this context, we created the Covid-on-the-Web project, which aims to assist the access, querying, and sense making of COVID-19 related literature by combining efforts from semantic web, natural language processing, and visualization fields. Particularly, in this paper, we present (i) an RDF dataset, a linked version of the "COVID-19 Open Research Dataset" (CORD-19), enriched via entity linking and argument mining, and (ii) the "Linked Data Visualizer" (LDViz), which assists the querying and visual exploration of the referred dataset. The LDViz tool assists 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.

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

1 Introduction

The COVID-19 pandemics motivated the scientific community from 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 datasets covering the publications about COVID-19 and related coronaviruses and diseases have been compiled to support the scientific community. Particularly, we focus on the *COVID-19 Open Research Dataset* (CORD-19) (Wang 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 relevant information from it without appropriate support. To help researchers to find publications of interests, we employ information visualization techniques to explore the dataset and identify relationships among publications that indicate those that are worthy of further examination.

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., semantic web, natural language processing, and visualization) to assist the exploration of the COVID-19 scientific literature. Through a series of interviews with our prospect 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 of the relationship between coronaviruses and other diseases (e.g., cancer), identifying the types of cancers that are likely to occur in COVID-19 victims, among others. Whilst 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

[1] https://www.inserm.fr/

[2] https://www.e-cancer.fr/

victims) and analyzing specific attributes (e.g., details about metabolic changes caused by COVID-19). Furthermore, the analysis of co-authorship is relevant to health research as it allows to assess collaboration trends and identify leading investigators and organizations (Fonseca 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 co-authorship networks, which allow to answer user queries such as “where are researches 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 dataset, a linked version of the CORD-19 corpus, enriched via entity linking and argument mining. Currently, the *Covid-on-the-Web RDF dataset* includes and enriches over 100,000 full-text scholarly articles from the 47^{th} version of the CORD-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 correspond to LDViz[3], a visualization tool that enables the exploration of the COVID-19 scientific literature from different perspectives, such as co-authorship, named entities co-occurrence and the relationship between claims and evidences within publications. We demonstrate the potential of LDViz to support the exploration of customizable SPARQL result sets extracted from the Covid-on-the-Web dataset to assist the resolution of different domain-related tasks.

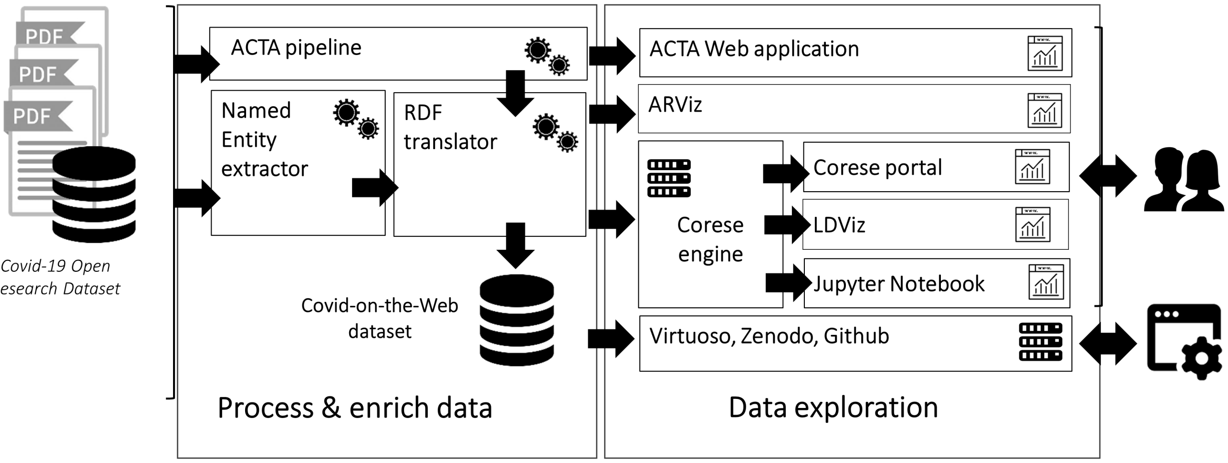
Although there have been previous contributions in exploring the CORD-19 corpus through entity linking approaches (e.g., Oniani et al., 2020; Reese et al., 2021), to the best of our knowledge, the Covid-on-the-Web dataset is the first to integrate NEs, arguments and PICO components into a single, coherent whole. Furthermore, we propose an unified pipeline (Figure 1) that facilitates the extraction and visualization of information from the CORD-19 corpus by continuously producing and publishing an enriched linked data

[3] Link for an illustration video of LDViz: https://youtu.be/Cn_IWQ7yVvE

knowledge graph. Also, our visualization approach differs from previous solutions to explore the COVID-19 scientific literature (e.g., Hope et al., 2020; Verspoor 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 help them to understand how they have arrived to a certain outcome.

Figure 1

Overview of the Covid-on-the-Web project: pipeline, resources, services and applications.



The remaining of this paper is organized as follows. Section 2 presents previous data mining and visualization approaches to explore the CORD-19 corpus. Section 3 describes the extraction pipeline to process the CORD-19 corpus and generate the RDF dataset and presents the characteristics of the dataset 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 dataset. Finally, section 7 concludes this paper.

2 Related Work

Since March of 2020, when the CORD-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 works related to the exploration of the CORD-19 dataset in terms of data enrichment and visualization.

2.1 Data Enrichment

Entity linking is usually the first approach for processing or enriching a dataset, which we can observe in several initiatives throughout the literature, such as: the CORD-19-on-FHIR (Oniani et al., 2020) project, which transforms the CORD-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 et al., 2020) project, which seeks the discovery of drug repurposing hypothesis 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 NEs linked to Wikidata concepts. Furthermore, these solutions are mostly focused on biomedical ontologies, resulting in NEs strongly related to genes, proteins, drugs, diseases, phenotypes and publications, while we extend the scope of ontologies to include DBpedia and Wikidata, resulting in named entities that go beyond the biological domain to extend the scope of analysis. Furthermore, we integrate argumentation structures and named entities in a coherent dataset.

2.2 Visualization Approaches

The Covid19-PubAnnotation[4] repository gathers text annotations regarding the CORD-19 corpus and other COVID-19 datasets. The annotations are recovered from multiple sources and 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 COVID-19 scientific literature and supports browsing through networks of biomedical concepts and research groups. It automatically extracts textual and co-authorship 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 co-authors 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 CORD-19 corpus through three different views: a sankey diagram displays the relationship between PICO concepts and allows 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 et al., 2020) interface uses semantic visualization to explore the publications within the CORD-19 and other COVID-19 datasets. It provides three

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

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

Bras 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 CORD-19 dataset. The system provides a cluster visualization displaying all resources in the dataset, where the user can select a resource to explore its related topics, descriptions, trend analysis, and documents.

The CovidExplorer (Ambavi et al., 2020) is a multi-faceted 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 CORD-19 dataset. It comprises a query interface that supports keyword-based search of authors, papers (title), and full-text papers; and a named entity recognition system which computes indicators of first mention of entities, popular co-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 yet a spatio-temporal 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, 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 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 datasets via SPARQL SELECT queries applied to any SPARQL endpoint (illustrated here via the Covid-on-the-Web dataset), so that they can explore multiple aspects of RDF datasets 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 dataset, 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. named entities, arguments, etc) we perform in the Covid-on-the-Web dataset, 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" Dataset

In this section, we describe the *Covid-on-the-Web* dataset which we produced from processing and analyzing the CORD-19 corpus. The dataset cohesively integrates the results of two mining processes: (1) a named entities (NE) extraction and linking that define the links between the CORD-19 articles and major public datasets of the Web of Data, and (2) an extraction of argumentative components discovered in the articles. These are both represented as RDF knowledge graphs described hereafter.

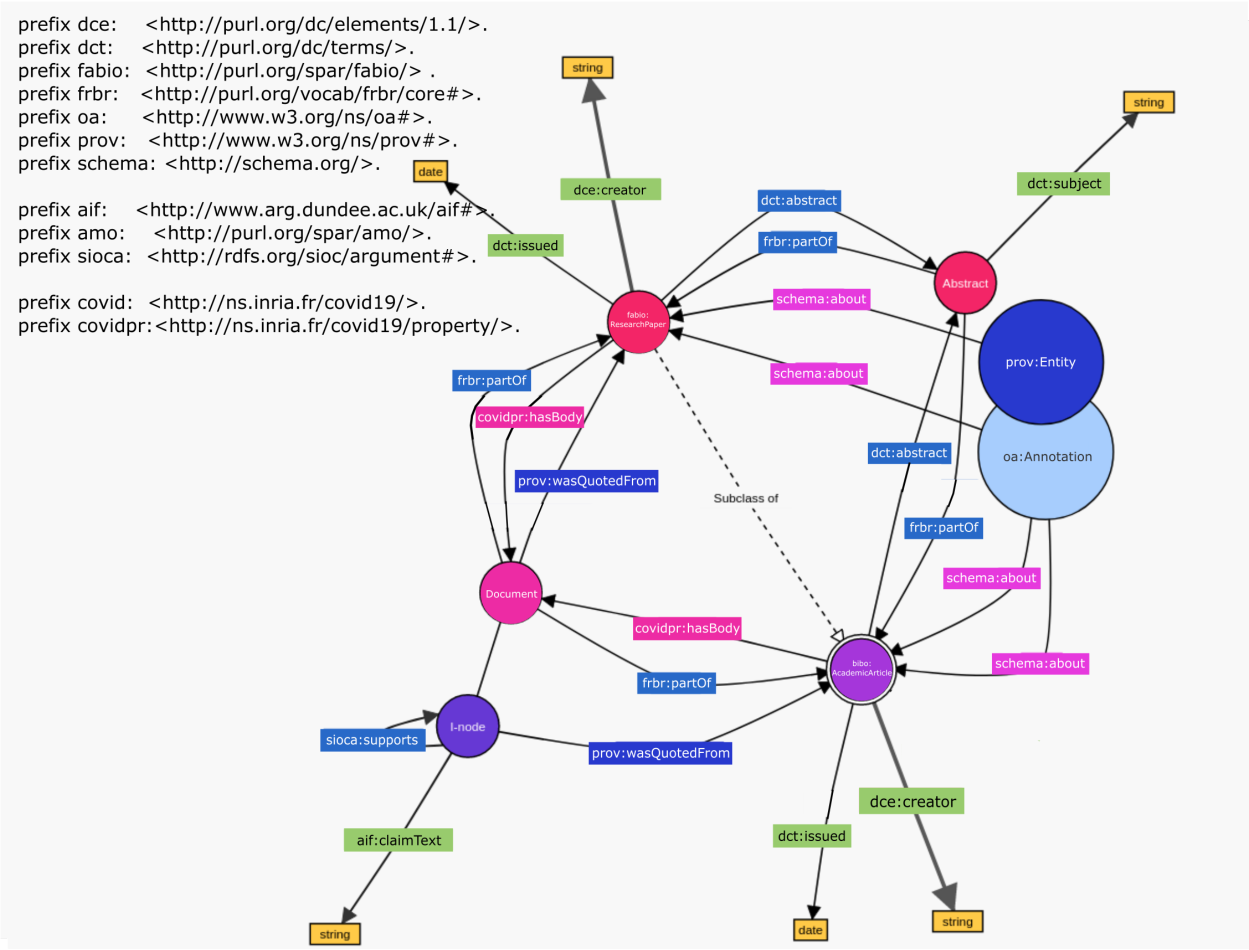
3.1 The CORD-19 Named Entities Knowledge Graph

The *CORD-19 Named Entities Knowledge Graph* (CORD19-NEKG) represents NEs identified and disambiguated in the articles of the CORD-19 corpus using three tools: DBpedia Spotlight (Daiber et al., 2013) to disambiguate NEs against DBpedia entities; the

Entity-fishing[5] tool to disabiguate NEs against Wikidata entities; and NCBO BioPortal Annotator (Jonquet et al., 2009) to disambiguate NEs against entities found in BioPortal's ontologies.

Figure 2

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



```
prefix dce:    <http://purl.org/dc/elements/1.1/>.
prefix dct:    <http://purl.org/dc/terms/>.
prefix fabio:  <http://purl.org/spar/fabio/> .
prefix frbr:   <http://purl.org/vocab/frbr/core#>.
prefix oa:     <http://www.w3.org/ns/oa#>.
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#>.

prefix covid:  <http://ns.inria.fr/covid19/>.
prefix covidpr:<http://ns.inria.fr/covid19/property/>.
```

CORD19-NEKG uses common, well-adopted terminological resources to represent

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

articles and NEs in RDF. We use DCMI[6], FaBiO[7], the Bibliographic Ontology[8], FOAF[9], and Schema.org[10] to represent article metadata such as the title, authors and DOI, and the Web Annotation Vocabulary[11] and Provenance Ontology[12] 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, which full description together with examples is available in the project's Github repository.[13]

3.2 The CORD-19 Argumentative Knowledge Graph

The *ACTA* (Argumentative Clinical Trial Analysis) (Mayer et al., 2019) tool was originally designed to help clinicians make decisions in evidence-based medicine by automatically extracting argumentative components and PICO elements[14] 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

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

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

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

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

[10] https://schema.org/

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

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

[13] https://github.com/Wimmics/covidontheweb

[14] 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).

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 et al., 2019). While 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, since arguments can be extracted from abstracts not necessarily dealing with clinical trials and PICO elements detection can be generalized to every biomedical article, we re-purposed ACTA to also annotate the articles from the CORD-19 corpus. Thus, we analyzed every abstract and translated the result into RDF to create the *CORD-19 Argumentative Knowledge Graph* (CORD19-AKG), which represent the argumentative components through the Argument Model Ontology (AMO)[15], the SIOC Argumentation Module (SIOCA)[16] and the Argument Interchange Format[17]. Further, the PICO elements are described as annotations of the argumentative components in a similar way to the NEs and disambiguated against UMLS concepts and semantic types.

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

The Covid-on-the-Web dataset has a DOI and can be downloaded from Zenodo[18]. It can also be queried through our public SPARQL endpoint[19]. The RDF dataset embeds detailed metadata 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 et al., 2020.

[15] http://purl.org/spar/amo/

[16] http://rdfs.org/sioc/argument#

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

[18] Dataset DOI: 10.5281/zenodo.4247134. Download page: https://doi.org/10.5281/zenodo.4247134

[19] SPARQL Enpoint https://covidontheweb.inria.fr/sparql

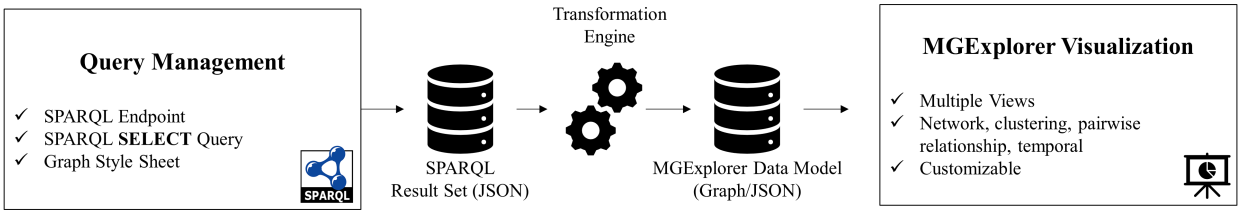
[20] http://ns.inria.fr/covid19/covidontheweb-1-2

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 datasets 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.

Figure 3

Linked Data Visualizer architecture overview. (a) Query Management Interface. (b) Transformation engine. (c) Visualization Interface supported by MGExplorer visualization tool.



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 dataset through a set of use case scenarios presented in Section 5.

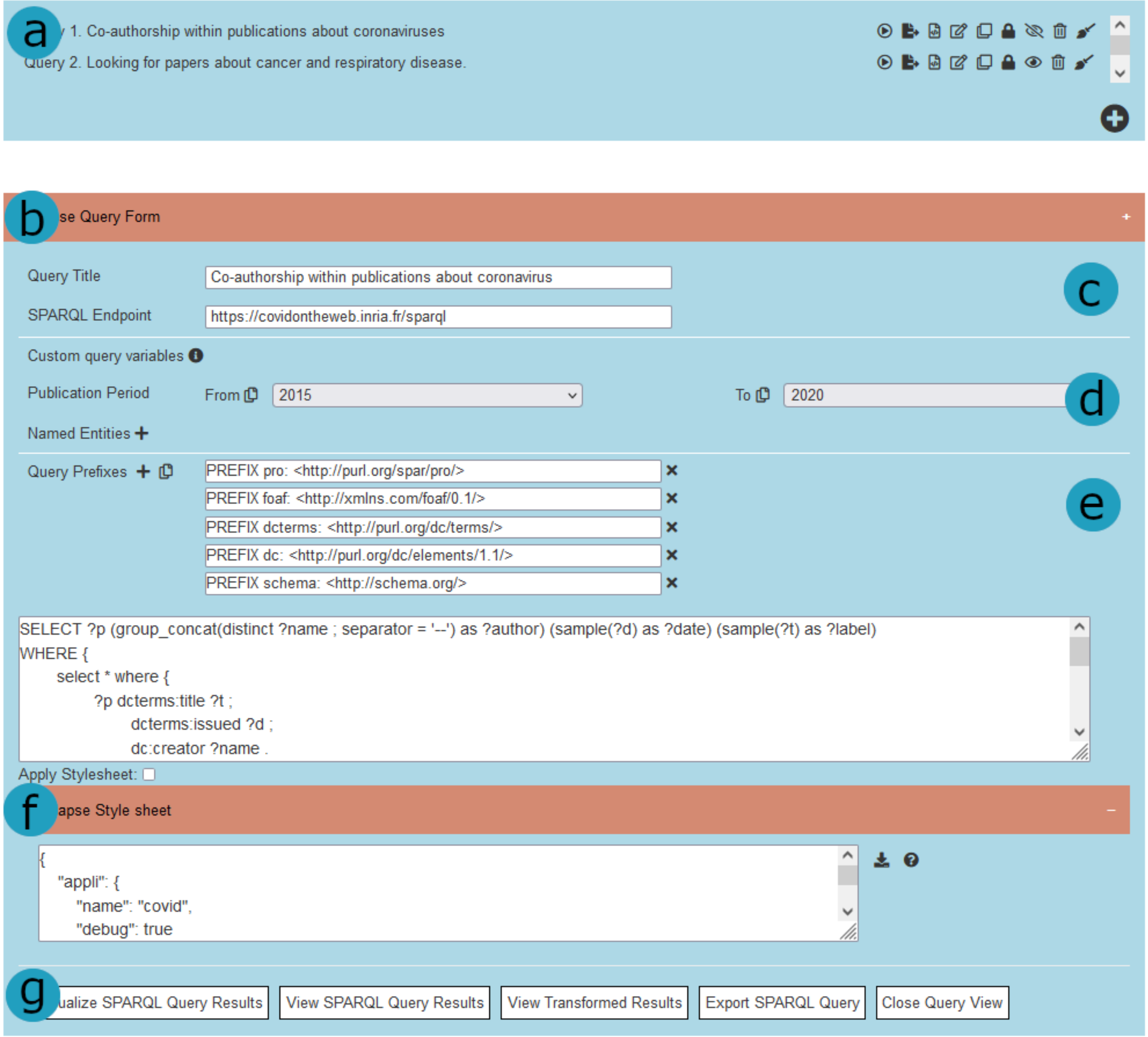
4.1 Query Management Interface

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, while Figure 4b-e depicts the interface areas enabling the edition 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 of 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 dataset. This flexibility allows to explore graphs with different topology (e.g., with nodes featuring publications, authors, named entities, etc.). 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 named entities) related by a particular document identified by a variable `?p`. An alternative to `?s`, `?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 human's perceptual and cognitive limits towards visualizations, 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 more relevant ones based on the number of bindings and classifies the remaining values as "Other". The `?label` variable allows to provide a description of 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).

Figure 4

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.



y 1. Co-authorship within publications about coronaviruses

Query 2. Looking for papers about cancer and respiratory disease.

Query Title: Co-authorship within publications about coronavirus

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

Publication Period: From 2015 To 2020

```
PREFIX pro: <http://purl.org/spar/pro/>
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 .
```

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

When dealing with a new dataset, researchers often have to debug and test multiple queries to discover the contents of the dataset. For the purpose of easing 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 publication period and named entities of interest (see Listing 1 for an example).

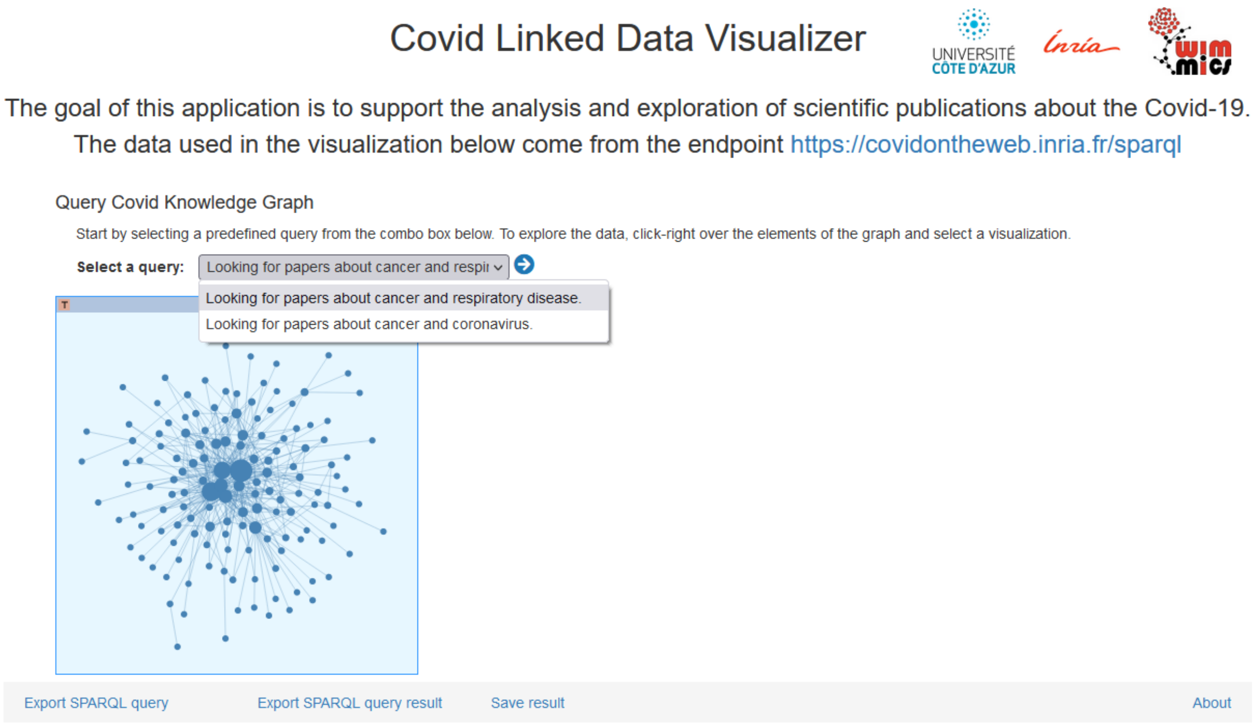
A Graph Style Sheet language (GSS) serves to transform the default node-link diagrammatic representation through the declarative specification of visibility, layout and styling rules applied to its nodes and arcs (Pietriga, 2006). Based on this concept, we associate each query to a GSS that the user 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 et al., 2012) is a RDF processor that enables among others 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 Subsection 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 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 intends to help expert users (developers and data scientists) to create suitable SPARQL

Figure 5

Public vitrine of Covid-19 Linked Data Visualizer.



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

UNIVERSITÉ CÔTE D'AZUR
Inria
wimmics
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

queries for exploring the dataset. However, expert users such as biomedical researchers do not need to know SPARQL for visualizing and interacting with the results of queries. Indeed, they may benefit of a public vitrine[21] 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 of those queries by means of an information visualization interface.

[21] Accessible at http://covid19.i3s.unice.fr:8080/

4.2 Visualization Interface

As mentioned earlier, LDViz uses the MGExplorer (**M**ultidimensional **G**raph **Explorer**) (Menin, Cava, et al., 2021) graphic library to support the visual exploration of the Covid-on-the-Web dataset. 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, 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:

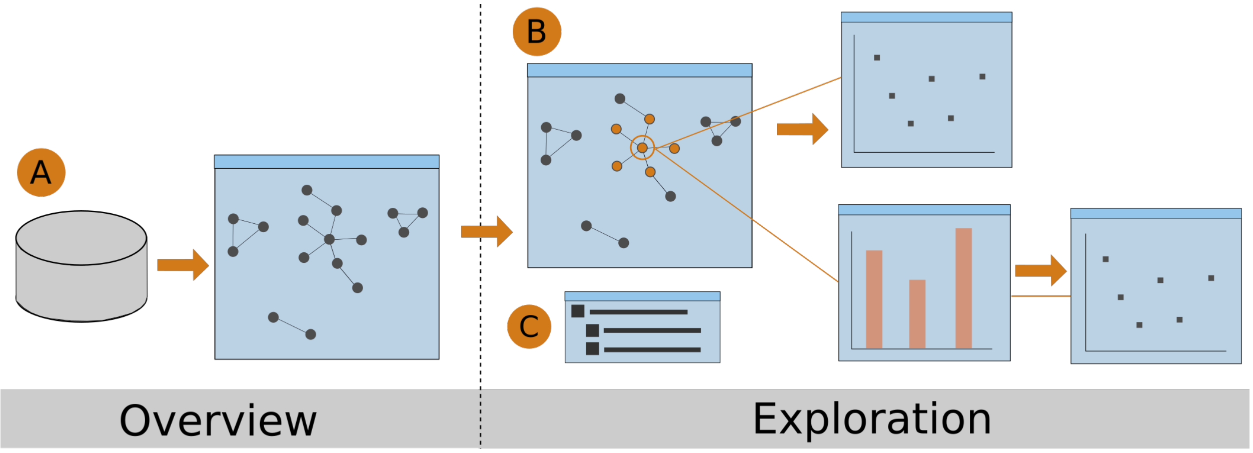
1. the *overview phase*, which consists of visualizing the network defined by the SPARQL query results through a node-link diagram (see description below). This visualization allows the user to get an overall understanding of the clusters within the data; and
2. the *exploratory phase*, where the user 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 (1) the comparison of two or more different subsets of data through a particular perspective provided by a particular visualization, and (2) the comparison of different perspectives of the same subset of data using multiple, complementary visualization techniques. Particularly, 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 provide 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 co-authorship networks or co-occurrence of named entities.

Figure 6

Overview of MGExplorer. (a) The node-link diagram provides an overview of the dataset. (b) Filtering operations enable further exploration of items/subsets of interest through different visualization techniques. (c) A history panel records users' actions throughout the exploration process. Image retrieved from (Menin, Cava, et al., 2021).



- The **ClusterVis** technique (Cava et al., 2017) enables the inspection of clusters and data attributes (e.g., publication type) within the subset of items (e.g., authors or named entities). The visualization has a multi-ring 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 belong to the same sub-cluster 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 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. In order 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.

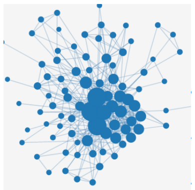
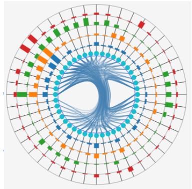
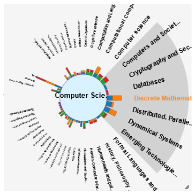
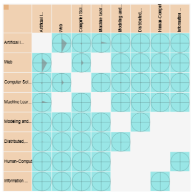
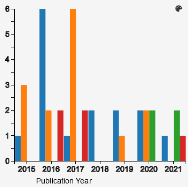
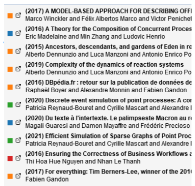
- The **GlyphMatrix** technique (Cava & Freitas, 2013) features a matrix where rows and columns represent data items (e.g., authors or named entities), and the intersection cell between each pair of items contains a glyph encoding the data attributes describing that relationship. The default glyph is 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 cell to make the glyph larger and more visible through a tooltip feature. This visualization technique could be seen as a combo 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, while the y-axis encodes the counting of publications. The data is 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 co-authored by two or more authors or the publications where two or more named entities co-occur, according to 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).

Table 1

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

Node-link Diagram	ClusterVis	IRIS	GlyphMatrix	Bar chart	Listing
					(2017) A MODEL-BASED APPROACH FOR DESCRIBING OFFI Marco Winckler and Félix Albertos Marco and Victor Penichet (2016) A Theory for the Composition of Concurrent Proces Eric Madelaine and Min Zhang and Ludovic Henrio (2015) Ancestors, descendants, and gardens of Eden in re Alberto Dennunzio and Luca Manzoni and Antonio Enrico Po (2019) Complexity of the dynamics of reaction systems Alberto Dennunzio and Luca Manzoni and Antonio Enrico Po (2016) DBpédia.fr : retour sur la publication de données de Raphaël Boyer and Alexandre Monnin and Fabien Gandon (2020) Discrete event simulation of point processes: A co Patricia Reynaud-Bouret and Cyrille Mascart and Alexandre I (2020) Du texte à l'intertexte. Le palimpseste Macron au r Magali Guaresi and Damon Mayaffre and Frédéric Precioso (2021) Efficient Simulation of Sparse Graphs of Point Proc Patricia Reynaud-Bouret and Cyrille Mascart and Alexandre I (2016) Ensuring the Correctness of Business Workflows a Thi Hoa Hue Nguyen and Nhan Le Thanh (2017) For everything: Tim Berners-Lee, winner of the 201 Fabien Gandon
network	clusters	pairwise		distribution	listing

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 6-C) 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 recover 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 dataset. 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 multi-perspective exploration of the dataset.

5.1 Scenario 1. Clusters Visualization

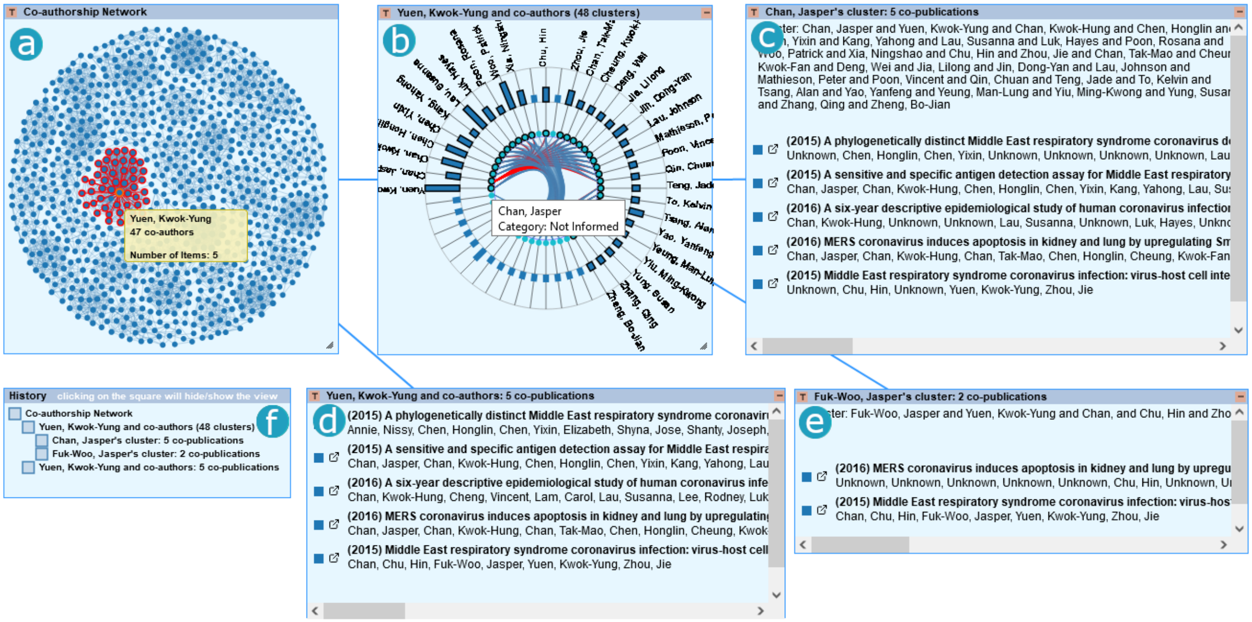
Based on the premise that COVID-19 has increased the collaboration between researchers from diverse disciplines around the world (Naujokaitytė, 2021), a biomedical researcher from INCa was interested on searching for information about existing collaborations on the theme of the relationship between COVID-19 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 co-authorship networks.

We use a subset of data describing the co-authorship 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 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 co-authors and related information to their co-publications. As mentioned earlier, the MGExplorer visualization interface self-starts with an overview of co-authorship 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 sub-graph related to the author Yuen, Kwok-Yung (Figure 7a), who will be our author of interest for this exploration. We hover

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 sub-clusters within the set of co-authors and their co-publications. (c)-(e) The views depict the publications produced within each sub-cluster. (e) The total publications of the author of interest. (f) The history shows which charts were opened, their order and inner dependencies.



Chan, Jasper's cluster: 5 co-publications

ster: Chan, Jasper and Yuen, Kwok-Yung and Chan, Kwok-Hung and Chen, Honglin an
n, Yixin and Kang, Yahong and Lau, Susanna and Luk, Hayes and Poon, Rosana and
Woo, Patrick and Xia, Ningshao and Chu, Hin and Zhou, Jie and Chan, Tak-Mao and Cheun
Kwok-Fan and Deng, Wei and Jia, Lilong and Jin, Dong-Yan and Lau, Johnson and
Mathieson, Peter and Poon, Vincent and Qin, Chuan and Teng, Jade and To, Kelvin and
Tsang, Alan and Yao, Yanfeng and Yeung, Man-Lung and Yiu, Ming-Kwong and Yung, Susar
and Zhang, Qing and Zheng, Bo-Jian

- **(2015) A phylogenetically distinct Middle East respiratory syndrome coronavirus d**
  Unknown, Chen, Honglin, Chen, Yixin, Unknown, Unknown, Unknown, Unknown, Lau
- **(2015) A sensitive and specific antigen detection assay for Middle East respiratory**
  Chan, Jasper, Chan, Kwok-Hung, Chen, Honglin, Chen, Yixin, Kang, Yahong, Lau, Su
- **(2016) A six-year descriptive epidemiological study of human coronavirus infectio**
  Chan, Kwok-Hung, Unknown, Unknown, Lau, Susanna, Unknown, Luk, Hayes, Unkno
- **(2016) MERS coronavirus induces apoptosis in kidney and lung by upregulating Sm**
  Chan, Jasper, Chan, Kwok-Hung, Chan, Tak-Mao, Chen, Honglin, Cheung, Kwok-Fan
- **(2015) Middle East respiratory syndrome coronavirus infection: virus-host cell inte**
  Unknown, Chu, Hin, Unknown, Yuen, Kwok-Yung, Zhou, Jie

History clicking on the square will hide/show the view

- □ Co-authorship Network
  - □ Yuen, Kwok-Yung and co-authors (48 clusters)
    - □ Chan, Jasper's cluster: 5 co-publications
    - □ Fuk-Woo, Jasper's cluster: 2 co-publications
  - □ Yuen, Kwok-Yung and co-authors: 5 co-publications

Yuen, Kwok-Yung and co-authors: 5 co-publications

- **(2015) A phylogenetically distinct Middle East respiratory syndrome coronavir**
  Annie, Nissy, Chen, Honglin, Chen, Yixin, Elizabeth, Shyna, Jose, Shanty, Joseph,
- **(2015) A sensitive and specific antigen detection assay for Middle East respira**
  Chan, Jasper, Chan, Kwok-Hung, Chen, Honglin, Chen, Yixin, Kang, Yahong, Lau
- **(2016) A six-year descriptive epidemiological study of human coronavirus infe**
  Chan, Kwok-Hung, Cheng, Vincent, Lam, Carol, Lau, Susanna, Lee, Rodney, Luk
- **(2016) MERS coronavirus induces apoptosis in kidney and lung by upregulatin**
  Chan, Jasper, Chan, Kwok-Hung, Chan, Tak-Mao, Chen, Honglin, Cheung, Kwok-
- **(2015) Middle East respiratory syndrome coronavirus infection: virus-host cell**
  Chan, Chu, Hin, Fuk-Woo, Jasper, Yuen, Kwok-Yung, Zhou, Jie

Fuk-Woo, Jasper's cluster: 2 co-publications

ster: Fuk-Woo, Jasper and Yuen, Kwok-Yung and Chan, and Chu, Hin and Zho

- **(2016) MERS coronavirus induces apoptosis in kidney and lung by upregu**
  Unknown, Unknown, Unknown, Unknown, Unknown, Chu, Hin, Unknown, U
- **(2015) Middle East respiratory syndrome coronavirus infection: virus-hos**
  Chan, Chu, Hin, Fuk-Woo, Jasper, Yuen, Kwok-Yung, Zhou, Jie

over the node representing the author, where we observe that they have 47 co-authors, 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 ClusterViz view, where we can explore the different clusters within the subset of co-authors 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 co-authored together (Figure 7d-e). Finally, we could compare the contributions made within those clusters and the complete list of publications co-authored by our author of interest (Figure 7f), to understand the impact of these co-authorship relationships in terms of number and quality of publications they have together.

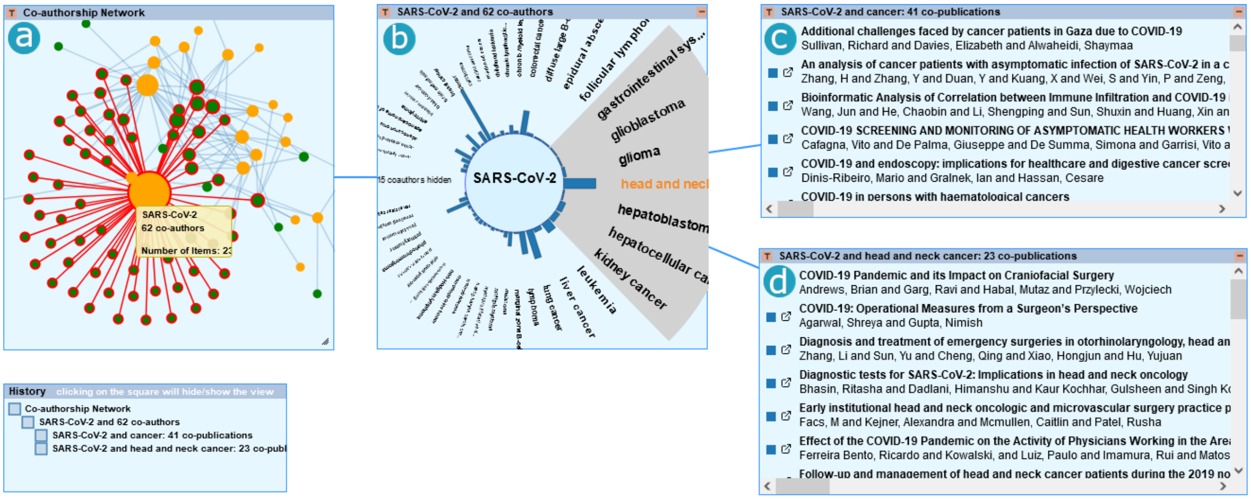
```
select ?p (group_concat(distinct ?name ; separator = '--') as ?author)
    (sample(?d) as ?date) (sample(?t) as ?label) where {
    select * where {
        ?doc dct:title ?t ; dct:issued ?d ; dce:creator ?name .
        filter contains(?t, "coronavirus")
        filter (?d >= "$beginYear-01-01"^^xsd:date) # $beginYear = 2015
        filter (?d <= "$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 co-authorship network within publications about "coronavirus" between 2015 and 2021.

5.2 Scenario 2. Customizing the Graph Topology

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.



SARS-CoV-2 and cancer: 41 co-publications

Additional challenges faced by cancer patients in Gaza due to COVID-19
Sullivan, Richard and Davies, Elizabeth and Alwaheidi, Shaymaa

An analysis of cancer patients with asymptomatic infection of SARS-CoV-2 in a c
Zhang, H and Zhang, Y and Duan, Y and Kuang, X and Wei, S and Yin, P and Zeng,

Bioinformatic Analysis of Correlation between Immune Infiltration and COVID-19 i
Wang, Jun and He, Chaobin and Li, Shengping and Sun, Shuxin and Huang, Xin an

**COVID-19 SCREENING AND MONITORING OF ASYMPTOMATIC HEALTH WORKERS **
Cafagna, Vito and De Palma, Giuseppe and De Summa, Simona and Garrisi, Vito a

COVID-19 and endoscopy: implications for healthcare and digestive cancer scre
Dinis-Ribeiro, Mario and Gralnek, Ian and Hassan, Cesare

COVID-19 in persons with haematological cancers

SARS-CoV-2 and head and neck cancer: 23 co-publications

COVID-19 Pandemic and its Impact on Craniofacial Surgery
Andrews, Brian and Garg, Ravi and Habal, Mutaz and Przylecki, Wojciech

COVID-19: Operational Measures from a Surgeon's Perspective
Agarwal, Shreya and Gupta, Nimish

Diagnosis and treatment of emergency surgeries in otorhinolaryngology, head an
Zhang, Li and Sun, Yu and Cheng, Qing and Xiao, Hongjun and Hu, Yujuan

Diagnostic tests for SARS-CoV-2: Implications in head and neck oncology
Bhasin, Ritasha and Dadlani, Himanshu and Kaur Kochhar, Gulsheen and Singh Ko

Early institutional head and neck oncologic and microvascular surgery practice p
Facs, M and Kejner, Alexandra and Mcmullen, Caitlin and Patel, Rusha

Effect of the COVID-19 Pandemic on the Activity of Physicians Working in the Area
Ferreira Bento, Ricardo and Kowalski, and Luiz, Paulo and Imamura, Rui and Matos

Follow-up and management of head and neck cancer patients during the 2019 no

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 co-authored). 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 the LDViz to solve this domain-related task. Using the query presented in Listing 3, we retrieve the RDF triples that correspond to the pattern `?s` → `?p` → `?o`, where `?s` and `?o` are, respectively, named entities related to (i.e., equal to, subclass of, or instance of) "cancer" and "coronavirus" named entities, and `?p` refers to the publications that contain these named entities 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 named entities instead of co-authors.

```
{"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
```

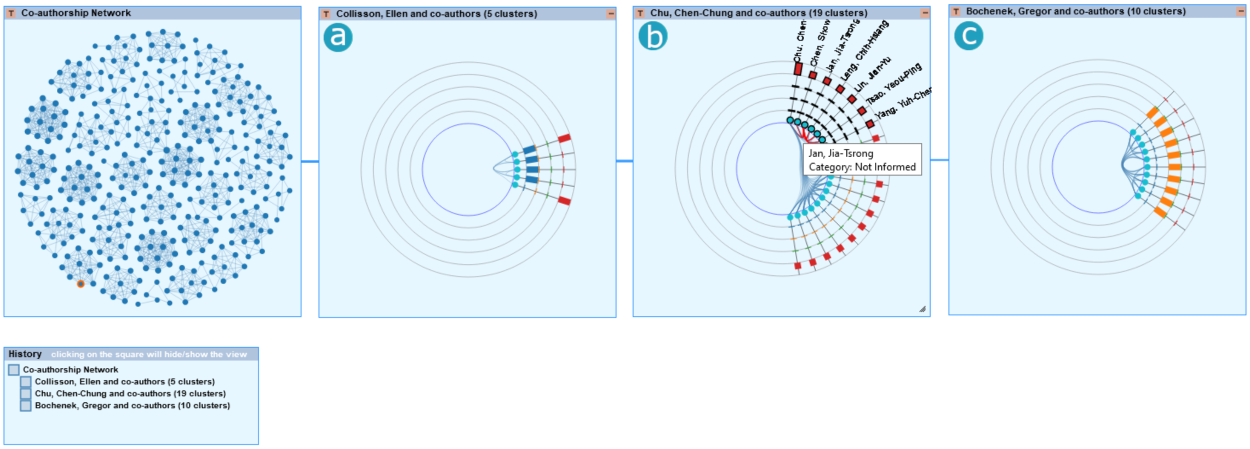
Listing 3: SPARQL query used in Use Case Scenarios 2 and 5 to retrieve the co-occurrence network within publications of named entities related to cancer and coronavirus.

Figure 8 depicts the exploratory path followed in this scenario to solve the above-described domain-related task. We explore a dataset that contains 452 RDF triples, which results in a graph with 94 nodes and 169 edges. Since in this dataset, we deal with two types of nodes, i.e., either related to “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 sub-graph originating from the SARS-CoV-2 named entity, which is associated to 62 types of cancer through 232 publications. We further explore the subset of data within this sub-graph by clicking-right 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 dataset, where the user can have more information about each document[22].

5.3 Scenario 3. Exploring Data Attributes

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 encode other subjects).


Co-authorship Network
Collisson, Ellen and co-authors (5 clusters)
a
Chu, Chen-Chung and co-authors (19 clusters)
b
Chu, Chen
Chen, Show
Jan, Jia-Tsrong
Leng, Chih-Hsiang
Lin, Jen-Yu
Tsao, Yeou-Ping
Yang, Yuh-Chen
Jan, Jia-Tsrong
Category: Not Informed
Bochenek, Gregor and co-authors (10 clusters)
c
History clicking on the square will hide/show the view
Co-authorship Network
Collisson, Ellen and co-authors (5 clusters)
Chu, Chen-Chung and co-authors (19 clusters)
Bochenek, Gregor and co-authors (10 clusters)

The previous exploration scenarios allow the user to see the relationship between co-authors or named entities, which can be characterized by the number of related publications. Thus, this scenario illustrates how we can use the LDViz to explore custom data attributes of a co-authorship network within coronavirus-related publications. In particular, we will use a dataset that describes 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 originate from the `schema:about` property, which refers to a set

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

of named entities that can be used to describe the research topic of the publication. The resulting dataset has 1,265 RDF triples, which were transformed in a graph with 356 nodes (authors) and 1,262 edges (co-publications). 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 co-authorship 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 co-authorship, while the publications co-authored by Chu, Chen-Chung (Figure 9b) refer to the “other” category of topics and are distributed throughout different clusters of co-authorship. Finally, we observe that the publication co-authored by Bocheneck, Gregor (Figure 9c), for instance, refers to the topic of “reverse transcriptase”.

```
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 co-authorship network within publications about “coronavirus” described by research subject.

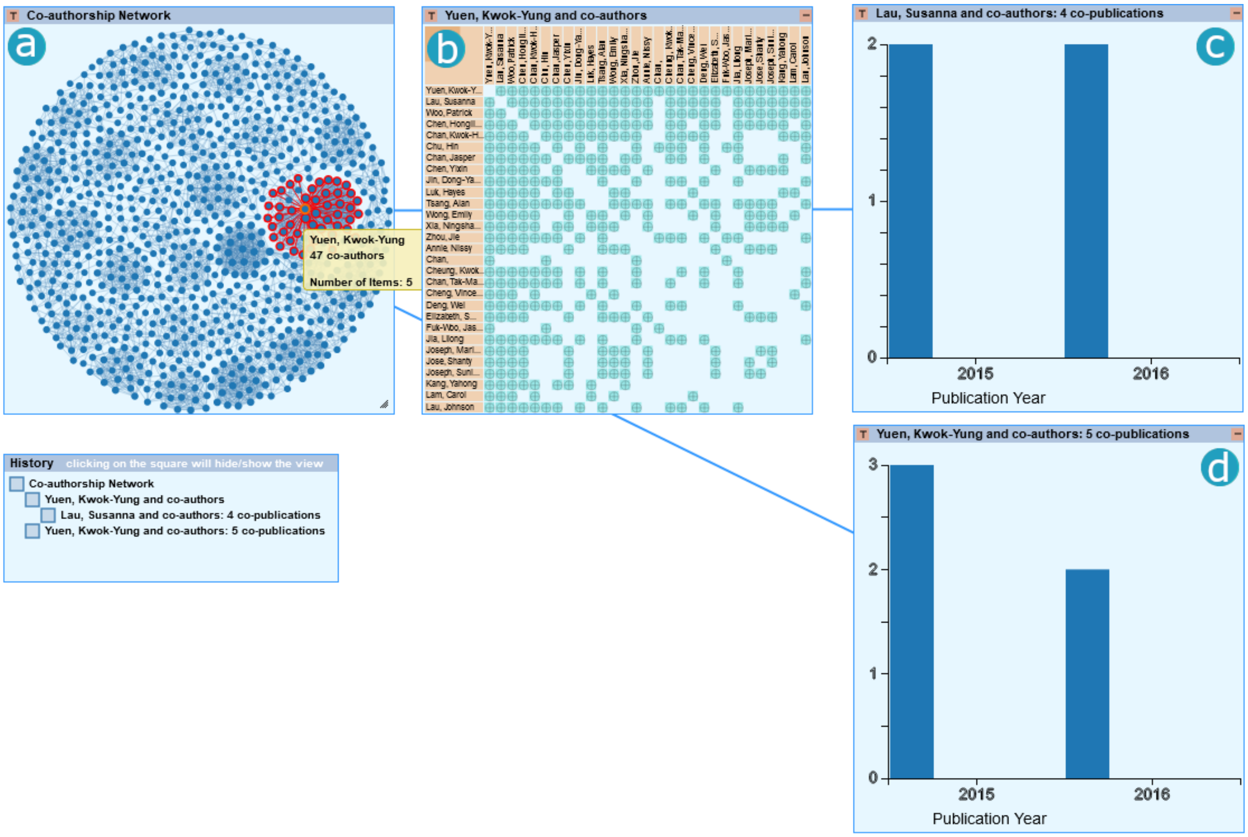
5.4 Scenario 4. Exploring the Temporal Aspect of Relationships

Studying the evolution over time of co-authors relationships or named entities co-occurrence could help understand when a collaboration between authors were stronger or when certain research topics were of higher 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 pandemics. Thus, in this scenario, we illustrate how one can use the LDViz interface to explore the temporal aspects of relationships, particularly co-authorship 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 co-authors, i.e., Yuen, Kwok-Yung (hereafter called author A) with 47 co-authors associated through five publications (Figure 10a). We further explore the relationship between author A and their co-authors through the GlyphMatrix visualization, which shows the types and number of co-publications between author A and every other co-author, as well as the co-publications among author A's co-authors. By ordering rows and columns by the number of co-publications, we can observe in the GlyphMatrix, that author A's most recurrent co-author is Lau, Susanna (herafter called author B) (Figure 10b), with whom they have 4 publications. Thus, to verify when these collaborations happened, we explore the temporal distribution of co-publications 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 2016. When comparing to the totality of co-publications related to author A (Figure 10d), we observe that four out of five publications are co-authored by author B which could indicate a strong collaboration between those authors in co-publications related to the coronavirus topic. We can also observe that this collaboration appear to have ended five years ago, since the dataset contains publications from 2015 to 2021.

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 co-author at the top-left cells, and we (c) explore the temporal distribution of their co-publications using the Histogram, which we compare with (d) the temporal distribution of publications co-authored by the author of interest.

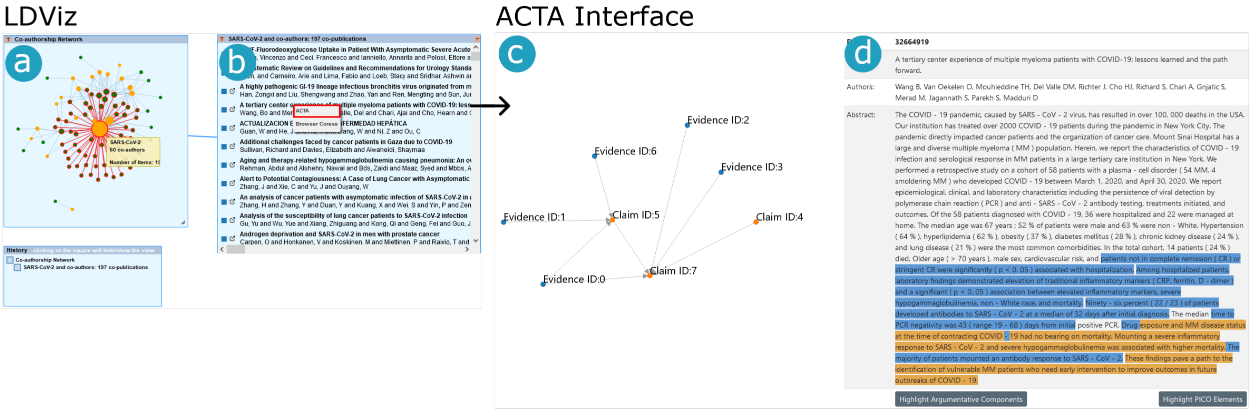


5.5 Scenario 5. Exploring Argumentation Graphs with the ACTA Interface

As mentioned earlier, the GSS feature allows the user to include external services in LDViz, such as a service that enable a 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., set of publications where named entities related to “cancer” and “coronavirus” co-occur) to illustrate how one can use the ACTA interface to visualize the

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.



SARS-CoV-2 and co-authors: 197 co-publications

- F-Fluorodeoxyglucose Uptake in Patient With Asymptomatic Severe Acute
  , Vincenzo and Ceci, Francesco and Ianniello, Annarita and Pelosi, Ettore a
- stematic Review on Guidelines and Recommendations for Urology Standa
  n, and Carneiro, Arie and Lima, Fabio and Loeb, Stacy and Sridhar, Ashwin ar
- A highly pathogenic GI-19 lineage infectious bronchitis virus originated from m
  Han, Zongxi and Liu, Shengwang and Zhao, Yan and Ren, Mengting and Sun, Jur
- A tertiary center e[...]ultiple myeloma patients with COVID-19: less
  Wang, Bo and Mer[...]alle, Del and Chari, Ajai and Cho, Hearn and C
- ACTUALIZACION E[...]NFERMEDAD HEPÁTICA
  Guan, W and He, J[...]iang, W and Ni, Z and Ou, C
- Additional challenges faced by cancer patients in Gaza due to COVID-19
  Sullivan, Richard and Davies, Elizabeth and Alwaheidi, Shaymaa
- Aging and therapy-related hypogammaglobulinemia causing pneumonia: An ov
  Rehman, Abdul and Alshehry, Nawal and Bds, Zaidi and Maaz, Syed and Mbbs, A
- Alert to Potential Contagiousness: A Case of Lung Cancer with Asymptomatic
  Zhang, J and Xie, C and Yu, J and Ouyang, W
- An analysis of cancer patients with asymptomatic infection of SARS-CoV-2 in a
  Zhang, H and Zhang, Y and Duan, Y and Kuang, X and Wei, S and Yin, P and Zen
- Analysis of the susceptibility of lung cancer patients to SARS-CoV-2 infection
  Gu, Yu and Wu, Yue and Xiang, Zhiguang and Kong, Qi and Geng, Fei and Guo, J
- Androgen deprivation and SARS-CoV-2 in men with prostate cancer
  Carpen, O and Honkanen, V and Koskinen, M and Miettinen, P and Raivio, T and

32664919

A tertiary center experience of multiple myeloma patients with COVID-19: lessons learned and the path forward.

Authors: Wang B, Van Oekelen O, Mouhieddine TH, Del Valle DM, Richter J, Cho HJ, Richard S, Chari A, Gnjatic S, Merad M, Jagannath S, Parekh S, Madduri D

Abstract: The COVID - 19 pandemic, caused by SARS - CoV - 2 virus, has resulted in over 100, 000 deaths in the USA. Our institution has treated over 2000 COVID - 19 patients during the pandemic in New York City. The pandemic directly impacted cancer patients and the organization of cancer care. Mount Sinai Hospital has a large and diverse multiple myeloma (MM) population. Herein, we report the characteristics of COVID - 19 infection and serological response in MM patients in a large tertiary care institution in New York. We performed a retrospective study on a cohort of 58 patients with a plasma - cell disorder (54 MM, 4 smoldering MM) who developed COVID - 19 between March 1, 2020, and April 30, 2020. We report epidemiological, clinical, and laboratory characteristics including the persistence of viral detection by polymerase chain reaction (PCR) and anti - SARS - CoV - 2 antibody testing, treatments initiated, and outcomes. Of the 58 patients diagnosed with COVID - 19, 36 were hospitalized and 22 were managed at home. The median age was 67 years ; 52 % of patients were male and 63 % were non - White. Hypertension (64 %), hyperlipidemia (62 %), obesity (37 %), diabetes mellitus (28 %), chronic kidney disease (24 %), and lung disease (21 %) were the most common comorbidities. In the total cohort, 14 patients (24 %) died. Older age (> 70 years), male sex, cardiovascular risk, and patients not in complete remission (CR) or stringent CR were significantly (p < 0. 05) associated with hospitalization. Among hospitalized patients, laboratory findings demonstrated elevation of traditional inflammatory markers (CRP, ferritin, D - dimer) and a significant (p < 0. 05) association between elevated inflammatory markers, severe hypogammaglobulinemia, non - White race, and mortality. Ninety - six percent (22 / 23) of patients developed antibodies to SARS - CoV - 2 at a median of 32 days after initial diagnosis. The median time to PCR negativity was 43 (range 19 - 68) days from initial positive PCR. Drug exposure and MM disease status at the time of contracting COVID - 19 had no bearing on mortality. Mounting a severe inflammatory response to SARS - CoV - 2 and severe hypogammaglobulinemia was associated with higher mortality. The majority of patients mounted an antibody response to SARS - CoV - 2. These findings pave a path to the identification of vulnerable MM patients who need early intervention to improve outcomes in future outbreaks of COVID - 19.

argumentative graph of a certain publication identified during the exploratory process. As one can see in Listing 2, the GSS form associate to the query contains an object called "services" that provides the redirection information for the ACTA interface (i.e., a call to "http://134.59.134.234:8081/analyseddocs?search="). The documents used in the Covid-on-the-Web dataset often originate from the PubMed archive[23], where each document has an 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 parameter.

Figure 11 depicts the exploratory path used in Scenario 5. As for Scenario 2, we identify the larger sub-graph 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

[23] https://pubmed.ncbi.nlm.nih.gov/

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 display the resulting argumentative graph with the relationships between claims and evidences, 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-AKG[24] dataset to explore claims and evidences graph related to one or more publications directly on LDViz by using a SPARQL query where `?s` and `?o` correspond to claims and evidences, while the `?p` variable correspond 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 CORD-19 dataset and enriched it via entity linking and argument mining. To the best of our knowledge, the Covid-on-the-Web dataset is the first public knowledge graph on the Web integrating publication metadata, named entities, arguments and PICO elements into a single, coherent whole. The openness aspect of our dataset and code should enable contributors to advance the current state of knowledge on this disease. Further, we believe the Covid-on-the-Web dataset could serve as a foundation for Semantic Web applications and benchmarking algorithms.

Moreover, we proposed a set of visualization interfaces to assist the exploration of

[24] http://ns.inria.fr/covid19/graph/acta

the Covid-on-the-Web dataset 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 the understanding of the sensemaking process. This visualization aims to help users understand the relationships within the results, e.g., 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 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 the Use Case Scenario 2 (see Subsection 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.

Additionally to our partners from 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. Particularly, 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 co-authorship 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 dataset 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 since 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 since 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 datasets in general, and in particular its suitability for analyzing the COVID-19 scientific literature and assisting 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 have 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 datasets (i.e., HAL open archive[25] and the Microsoft Academic Knowledge Graph[26], for which a set of predefined queries are available at http://covid19.i3s.unice.fr:8080/hal). The genericity of our approach enables the exploration of data from any SPARQL endpoint, such as DBpedia[27], from which we explored the ontology and RDF Schema information, as well as a co-starring relationship using movies information[28]. 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 endpoint[29], to directly visualize the resulting data. Furthermore, from a linked data perspective, one can use the Corese SPARQL

[25] https://data.archives-ouvertes.fr/doc/sparql

[26] https://makg.org/sparql

[27] http://fr.dbpedia.org/sparql

[28] Available at http://covid19.i3s.unice.fr:8080/ldviz

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

service[30] 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 the context of the 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 to the data and would create an alternative exploratory path to solve the same query. In this context, since visualization can help to recall, revisit, and reproduce the sensemaking process through visual representations of provenance data, MGExplorer visually represent the dependencies between views through line segments and uses the history panel to display exploratory actions hierarchically, keeping parenting and visualization information such as 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 a querying support for alternative datasets 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 datasets 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 make annotations throughout the process regarding the

[30] http://corese.inria.fr/sparql

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, extend and improve 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 names entities within publications through three complementary visualization techniques: a scatter plot, a chord diagram and an association graph[31]. The tool currently works separately with a pre-treated subset of data extracted from the Covid-on-the-Web dataset. However, the association mining algorithm can process any RDF dataset, which results could be then 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 on a completely integrated tool for extracting and exploring knowledge from scientific literature through various perspectives.

7 Conclusion

In this paper, we presented the dataset and software resources provided by the Covid-on-the-Web project, with particular focus on the visualization services proposed to

[31] Available at http://covid19.i3s.unice.fr:8080/arviz/

support the exploration of the COVID-19 scientific literature. Based on the needs of biomedical researchers, partners of the project, we designed and published a linked data knowledge graph describing the named entities mentioned in the articles of the CORD-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 dataset and the pipeline in meaningful ways to assist users needs.

Furthermore, we described and demonstrated the usage 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 dataset, supporting the resolution of diverse domain-related tasks.

Future works include evaluating our resources and services with participation of expert users in the biomedical domain in terms of usability and suitability to solve the domain-related tasks; developing of a querying feature that allows to dynamically import data into the exploratory process from external datasets, aiming to enrich the ongoing analysis and explore on-the-fly hypotheses; studying provenance information aiming on improving user experience and the visualization's effectiveness; and integrating new visualization services to extend the support for different domain-related tasks.

Acknowledgments

This work is partly funded by the French government labelled 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. 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.

References

Ambavi, H., Vaishnaw, 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*, 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.*

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. http://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*, 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*, 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. *Proc. 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*, 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*, 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://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-016-0104-5

Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., Hearst, M. A., & West, J. (2020). SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *arXiv preprint arXiv:2005.12668*.

Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N. T., Yao, Y., Rogers, C., Li, R., Liu, J., Singh, A., Schwabe, D., & Szekely, P. (2020). KGTK: A Toolkit for Large Knowledge Graph Manipulation and Analysis. *The Semantic Web – ISWC 2020*, 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*, 56. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041576/

Leng, J. (2011). *Handbook of research on computational science and engineering: Theory and practice* (Vol. 2). IGI Global.

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)*, 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.*

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*.

Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., Corby, O., Gazzotti, R., Giboin, A., Marro, S., Mayer, T., Simon, M., Villata, S., & 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 International Publishing. https://doi.org/10.1007/978-3-030-62466-8_19

Naujokaitytė, G. (2021). COVID-19 triggered unprecedented collaboration in research [Accessed on 6 July 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

Pietriga, E. (2006). Semantic web data visualization with graph style sheets. *Proceedings of the 2006 ACM symposium on Software visualization*, 177–178. https://doi.org/10.1145/1148493.1148532

Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., Shefchek, K. A., Good, B. M., Balhoff, J. P., Fontana, T., et al. (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

Sukla, A., Naskar, A., Goel, T., Sangwan, S., Rai, A., Shakir, M., Verma, I., Dasgupta, T., & 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.*

Verspoor, K., Šuster, S., Otmakhova, Y., Mendis, S., Zhai, Z., Fang, B., Lau, J. H., Baldwin, T., Yepes, A. J., & Martinez, D. (2020). COVID-SEE: Scientific Evidence Explorer for COVID-19 related research. *arXiv preprint arXiv:2008.07880.*

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). CORD-19: The Covid-19 Open Research Dataset. *ArXiv, abs/2004.10706.*

8 Author contributions (Contributor Roles Taxonomy Project – CRediT)

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

Raphael 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, Writing – original draft, Writing – review & editing, Supervision, Methodology, Formal analysis

The authors have no competing interests regarding this work.