



HAL
open science

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.

► To cite this version:

Aline Menin, Franck Michel, Fabien Gandon, Raphaël Gazzotti, Elena Cabrio, et al.. Covid-on-the-Web: Exploring the COVID-19 Scientific Literature through Visualization of Linked Data from Entity and Argument Mining. Quantitative Science Studies, 2021, 10.1162/qss_a_00164 . hal-03404580

HAL Id: hal-03404580

<https://hal.science/hal-03404580>

Submitted on 26 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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


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


6 **Author Note**

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

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

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

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

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

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

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

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

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

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

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

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

Abstract

19

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

36

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

37 1 Introduction

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

50 In collaboration with biomedical researchers from the French Institute of Medical
51 Research (Inserm)¹ and the French National Cancer Institute (INCa)², we created the
52 Covid-on-the-Web project, which gathers expertise from various research fields (i.e.,
53 semantic web, natural language processing, and visualization) to assist the exploration of
54 the COVID-19 scientific literature. Through a series of interviews with our prospect users,
55 we could identify a set of meaningful use case scenarios, such as determining the right
56 amount of certain substances in the patients' organism using baseline information collected
57 from scientific articles, analyzing clinical trials to make evidence-based decisions, studying
58 of the relationship between coronaviruses and other diseases (e.g., cancer), identifying the
59 types of cancers that are likely to occur in COVID-19 victims, among others. Whilst some
60 scenarios require exploring the relationship between components (e.g., cancer and
61 coronavirus), others require representing trends (e.g., probability of cancer in COVID-19

¹ <https://www.inserm.fr/>

² <https://www.e-cancer.fr/>

62 victims) and analyzing specific attributes (e.g., details about metabolic changes caused by
63 COVID-19). Furthermore, the analysis of co-authorship is relevant to health research as it
64 allows to assess collaboration trends and identify leading investigators and
65 organizations (Fonseca et al., 2016). In this paper, we focus on using visualization to assist
66 the resolution of user queries based on the relationship between components and
67 co-authorship networks, which allow to answer user queries such as “where are researches
68 in a particular topic being performed?”.

69 We present two contributions of the Covid-on-the-Web project to the exploration of
70 COVID-19 scientific literature. The first contribution refers to the Covid-on-the-Web RDF
71 dataset, a linked version of the CORD-19 corpus, enriched via entity linking and argument
72 mining. Currently, the *Covid-on-the-Web RDF dataset* includes and enriches over 100,000
73 full-text scholarly articles from the 47th version of the CORD-19 corpus, which corresponds
74 to 1.3 billion RDF triples describing the articles’ metadata, an argumentation and a named
75 entities (NE) knowledge graph. The second contribution correspond to LDViz³, a
76 visualization tool that enables the exploration of the COVID-19 scientific literature from
77 different perspectives, such as co-authorship, named entities co-occurrence and the
78 relationship between claims and evidences within publications. We demonstrate the
79 potential of LDViz to support the exploration of customizable SPARQL result sets
80 extracted from the Covid-on-the-Web dataset to assist the resolution of different
81 domain-related tasks.

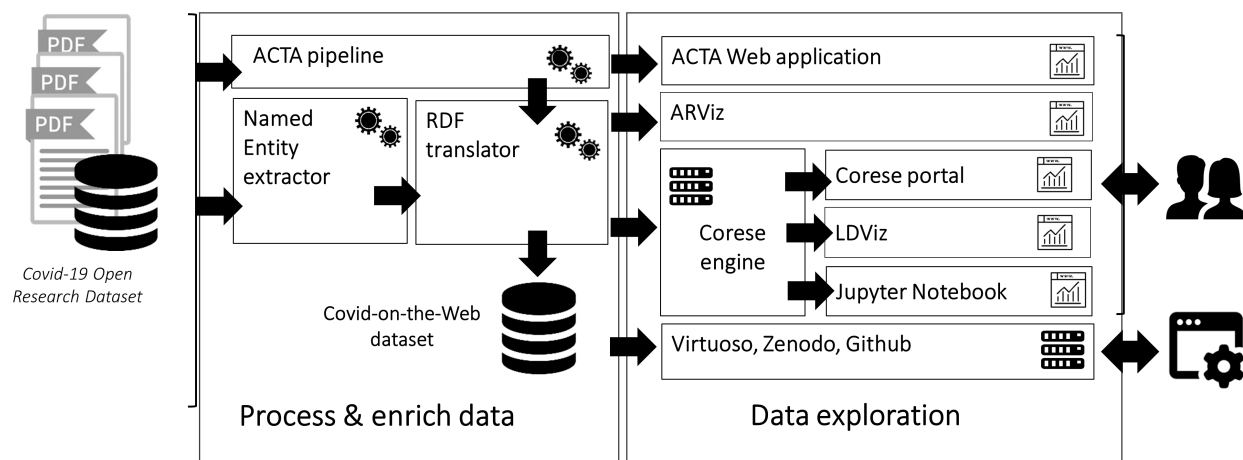
82 Although there have been previous contributions in exploring the CORD-19 corpus
83 through entity linking approaches (e.g., Oniani et al., 2020; Reese et al., 2021), to the best
84 of our knowledge, the Covid-on-the-Web dataset is the first to integrate NEs, arguments
85 and PICO components into a single, coherent whole. Furthermore, we propose an unified
86 pipeline (Figure 1) that facilitates the extraction and visualization of information from the
87 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

88 knowledge graph. Also, our visualization approach differs from previous solutions to
 89 explore the COVID-19 scientific literature (e.g., Hope et al., 2020; Verspoor et al., 2020), by
 90 supporting the exploration of meaningful subsets of data suitable to users' needs through
 91 the definition of custom SPARQL SELECT queries and via multiple, complementary
 92 visualization techniques; and by allowing the user to trace back their exploratory path,
 93 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.



94 The remaining of this paper is organized as follows. Section 2 presents previous
 95 data mining and visualization approaches to explore the COVID-19 corpus. Section 3
 96 describes the extraction pipeline to process the COVID-19 corpus and generate the RDF
 97 dataset and presents the characteristics of the dataset and the available services to exploit
 98 it. Section 4 describes LDViz, which usage and exploration potentials are demonstrated
 99 through use case scenarios in Section 5. Section 6 discusses future applications and
 100 potential impact of the dataset. Finally, section 7 concludes this paper.

101 **2 Related Work**

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

107 **2.1 Data Enrichment**

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

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

124 **2.2 Visualization Approaches**

125 The Covid19-PubAnnotation⁴ repository gathers text annotations regarding the
126 CORD-19 corpus and other COVID-19 datasets. The annotations are recovered from
127 multiple sources and aligned to the canonical text that is taken from PubMed and PMC
128 archives, which link annotations to each other. Furthermore, the platform provides simple
129 visualization that allows one to view the annotations directly on the text and further
130 explore them through interaction.

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

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

148 The SemViz (Tu et al., 2020) interface uses semantic visualization to explore the
149 publications within the CORD-19 and other COVID-19 datasets. It provides three

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

150 visualization techniques: a tag cloud gives an overall view of the most important concepts
151 within the data; a heat map represents a pairwise relationship between selected entities in
152 the article abstracts and journal names; and a data table is used to represent indexed
153 document data, such as sentences of biomedical relations and corresponding PubMed URLs
154 that link to the full article.

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

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

164 The CovidExplorer (Ambavi et al., 2020) is a multi-faceted AI-based search and
165 visualization engine that integrates search and recommendation, statistics, and social
166 media discussions to support the exploration of scientific articles from the CORD-19
167 dataset. It comprises a query interface that supports keyword-based search of authors,
168 papers (title), and full-text papers; and a named entity recognition system which computes
169 indicators of first mention of entities, popular co-mentioned entities, and year-wise
170 distribution of mention frequencies. These indicators are visualized through a timeline
171 chart and a sankey diagram, which shows the co-occurrence of entities within publications.
172 The system provides yet a spatio-temporal visualization of tweets regarding COVID-19.

173 Although we find several visualization tools to support either the exploration of
174 linked data in general or the COVID-19 scientific literature, as the ones presented above,
175 most of them support the exploration of raw data (i.e. the RDF graph, OWL or RDF
176 Schema), which is interesting for certain tasks such as exploring relevant concepts of an

177 application domain via ontology representation, inspecting RDF Graphs, and analyzing
178 instances based on their types/classes. Thus, we propose a flexible tool to enable users to
179 define meaningful datasets via SPARQL SELECT queries applied to any SPARQL
180 endpoint (illustrated here via the Covid-on-the-Web dataset), so that they can explore
181 multiple aspects of RDF datasets and the LOD Cloud. It also allows users to perform
182 exploratory searches using various complementary visualization techniques instantiated on
183 demand according to the task at hand, instead of a single visualization technique that
184 represents the whole dataset, restraining the analysis to a single view to the data. Our
185 approach is also based on a visualization concept that enables users to track their
186 exploratory path to help them to understand how they arrived to a certain outcome and to
187 allow them to explore alternative hypotheses generated on the fly through different
188 exploratory paths. Furthermore, the visualization together with the additional extractions
189 (i.e. named entities, arguments, etc) we perform in the Covid-on-the-Web dataset, enables
190 a deep and semantic-aware exploration of the topics and claims of the COVID scientific
191 corpus by leveraging the combination of semantic processing and exploratory search.

192 **3 The “Covid-on-the-Web” Dataset**

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

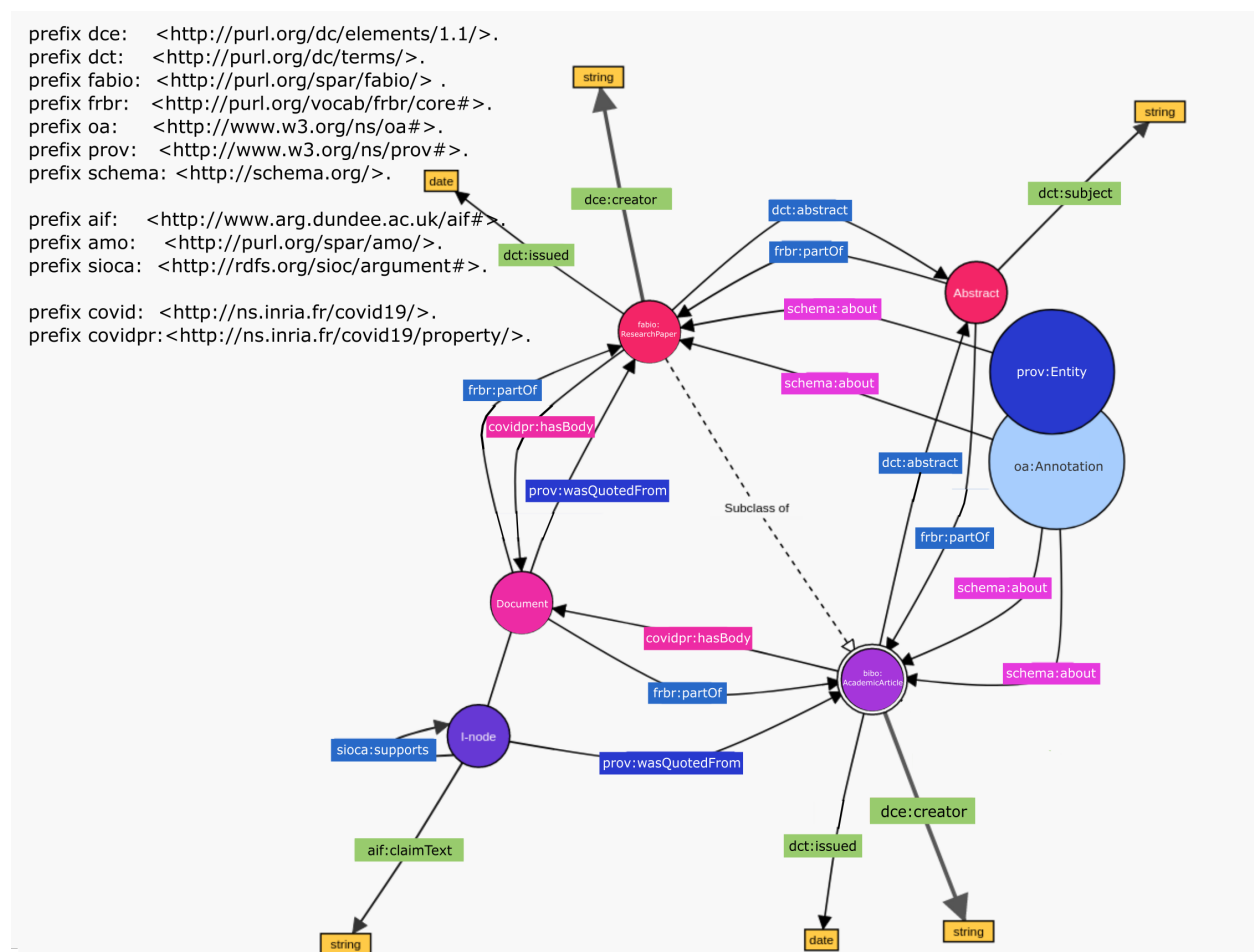
199 **3.1 The COVID-19 Named Entities Knowledge Graph**

200 The *COVID-19 Named Entities Knowledge Graph* (COVID19-NEKG) represents NEs
201 identified and disambiguated in the articles of the COVID-19 corpus using three tools:
202 DBpedia Spotlight (Daiber et al., 2013) to disambiguate NEs against DBpedia entities; the

203 Entity-fishing⁵ tool to disambiguate NEs against Wikidata entities; and NCBO BioPortal
 204 Annotator (Jonquet et al., 2009) to disambiguate NEs against entities found in BioPortal’s
 205 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).



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

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

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

215 3.2 The COVID-19 Argumentative Knowledge Graph

216 The *ACTA* (Argumentative Clinical Trial Analysis) (Mayer et al., 2019) tool was
217 originally designed to help clinicians make decisions in evidence-based medicine by
218 automatically extracting argumentative components and PICO elements¹⁴ from clinical
219 trials. Through multiple NLP steps, ACTA retrieves the argumentative components in the
220 trial and its PICO elements, classifies the components into *claim* (concluding statement)
221 and *evidence* (observation or measurement), and infers the relationship between the
222 components (i.e., *support* or *attack*). For instance, “a new treatment is considered more
223 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).

224 markers within the tested population (evidence)”.

225 The models used in ACTA are trained with SciBert, a language model for scientific
226 text, that has been shown to work on texts from different application domains (Beltagy
227 et al., 2019). While the content of articles might differ from clinical trials, the structure of
228 the abstracts is similar, including elements such as background, methods, results, and
229 conclusions. Thus, since arguments can be extracted from abstracts not necessarily dealing
230 with clinical trials and PICO elements detection can be generalized to every biomedical
231 article, we re-purposed ACTA to also annotate the articles from the COVID-19 corpus.
232 Thus, we analyzed every abstract and translated the result into RDF to create the
233 *COVID-19 Argumentative Knowledge Graph* (COVID19-AKG), which represent the
234 argumentative components through the Argument Model Ontology (AMO)¹⁵, the SIOC
235 Argumentation Module (SIOCA)¹⁶ and the Argument Interchange Format¹⁷. Further, the
236 PICO elements are described as annotations of the argumentative components in a similar
237 way to the NEs and disambiguated against UMLS concepts and semantic types.

238 3.3 Publishing and Querying Covid-on-the-Web Dataset

239 The Covid-on-the-Web dataset has a DOI and can be downloaded from Zenodo¹⁸. It
240 can also be queried through our public SPARQL endpoint¹⁹. The RDF dataset embeds
241 detailed metadata describing licensing, authorship, provenance, interlinking, and access
242 information, and the vocabularies used.²⁰ Additional information regarding reproducibility
243 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 Endpoint <https://covidontheweb.inria.fr/sparql>

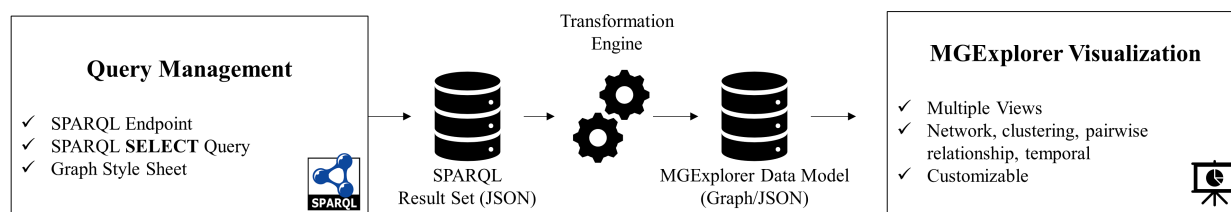
²⁰ <http://ns.inria.fr/covid19/covidontheweb-1-2>

244 **4 Linked Data Visualizer**

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

Figure 3

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



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

258 **4.1 Query Management Interface**

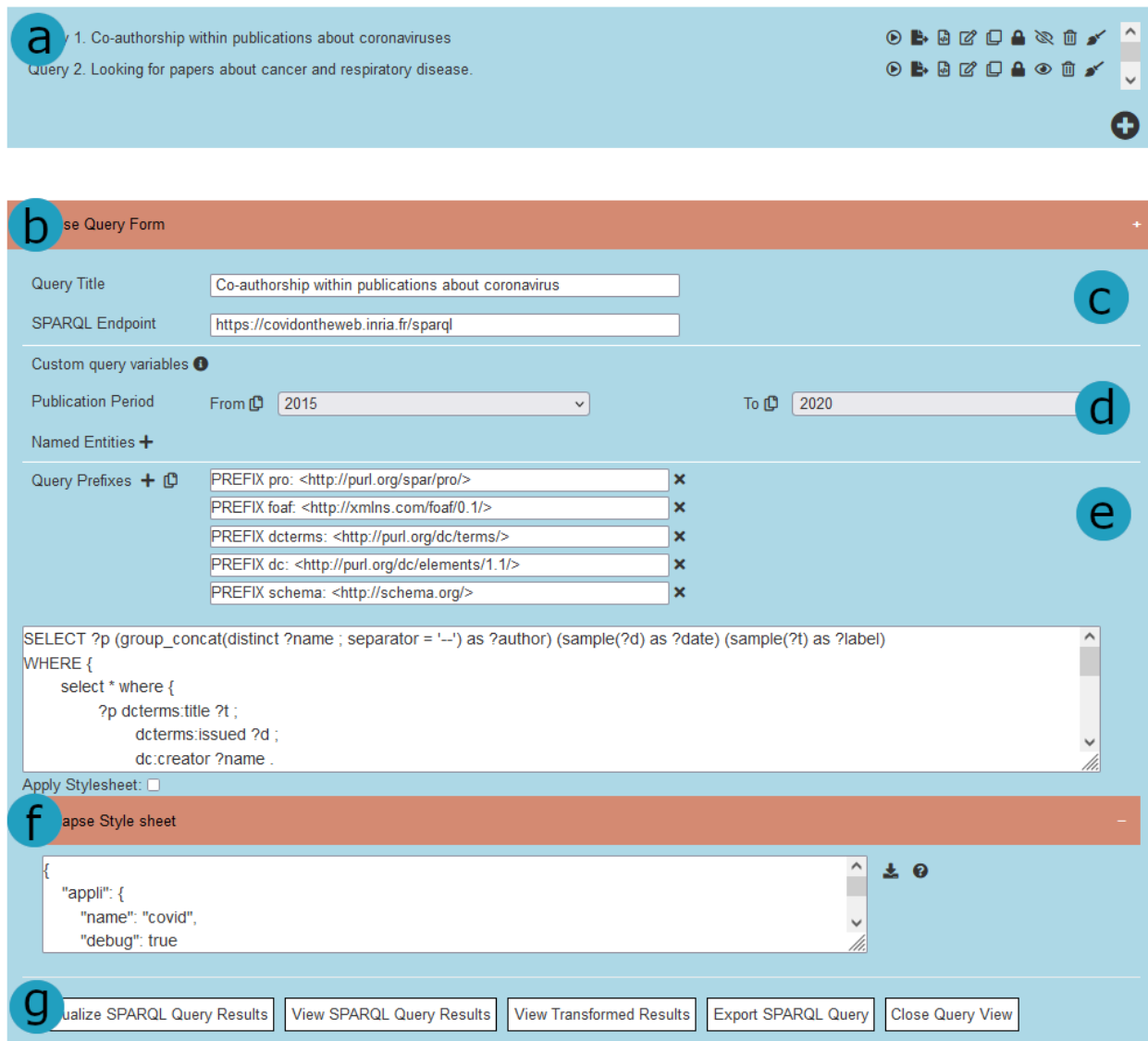
259 The query management interface (Figure 4) allows users to create and edit their
 260 own SPARQL queries. In Figure 4a, we can see the menu that lists and allows managing

261 predefined queries, while Figure 4b-e depicts the interface areas enabling the edition and
262 customization of queries. This interface also enables the preview and exporting of a query's
263 results (see Figure 4f). These can be visualized via the MGExplorer graphic library and/or
264 exported as JSON files containing either the results in the SPARQL JSON format or the
265 transformed results used as input of the visualization. The user can type the query in a
266 text area, which can include customizable parameters specified through HTML forms, such
267 as the publication date. Upon submitting a query, the results are processed by a
268 transformation engine that converts the SPARQL JSON format into the JSON format
269 expected by the graphic library.

270 The transformation engine is generic enough to support the exploration of different
271 variables of the dataset. This flexibility allows to explore graphs with different topology
272 (e.g., with nodes featuring publications, authors, named entities, etc.). In the context of
273 LDViz, this is made possible by using a SPARQL query that requires at least three
274 variables: `?s` and `?o`, which describe the nodes (e.g., authors or named entities) related by
275 a particular document identified by a variable `?p`. An alternative to `?s`, `?o` is the variable
276 `?author` which contains a list of authors. In addition to these variables, the system allows
277 three other reserved variables that serve to describe the edges (`?p`) of the output graph
278 visualization: `?type`, `?label`, and `?date`. The variable `?type` can be used to type the edges
279 of the output graph (e.g., by publication type). Due to human's perceptual and cognitive
280 limits towards visualizations, only a certain number of graphic elements can be drawn on
281 the screen. Thus, we allow the variable `?type` to be bound to only four different values
282 describing the edges. When it is bound to more than four distinct values in the SPARQL
283 query result, the system automatically determines the three more relevant ones based on
284 the number of bindings and classifies the remaining values as "Other". The `?label` variable
285 allows to provide a description of edges in natural language (e.g., the value of `rdfs:label`
286 properties describing resources). Finally, the `?date` variable is used to provide a visual
287 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.



288

When dealing with a new dataset, researchers often have to debug and test multiple

289

queries to discover the contents of the dataset. For the purpose of easing the customization

290 of queries and the use of the interface by the domain expert, we provide query templates
291 that allow one to interactively define the value of certain parameters, such as publication
292 period and named entities of interest (see Listing 1 for an example).

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


306 Although we demonstrate the usage of the querying and visualization interfaces for
307 exploring the Covid-on-the-Web dataset, LDViz can be used to query and visualize data
308 from any SPARQL endpoint. The querying form contains a field where the user enters the
309 endpoint URL, and the only requirement is that the query returns values for the
310 above-listed predefined set of variables. Hence, what we propose with LDViz is a generic
311 visualization tool for the Semantic Web of Linked Data.

312 As for any visualization, user queries must be translated to a query language that
313 recovers the necessary data from the database to solve the exploratory task. In this paper,
314 the user queries were identified during interviews with users from INCa and Inserm and
315 translated into SPARQL queries by data scientists. Thus, the query management interface
316 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 respiratory disease.

Looking for papers about cancer and coronavirus.



Export SPARQL query
Export SPARQL query result
Save result
About

317 queries for exploring the dataset. However, expert users such as biomedical researchers do
 318 not need to know SPARQL for visualizing and interacting with the results of queries.
 319 Indeed, they may benefit of a public vitrine²¹ simply by selecting a predefined query to
 320 explore the results with MGExplorer without having to deal with SPARQL expressions
 321 (Figure 5). The visibility of the predefined queries in the vitrine is settled when queries are
 322 created at the query management interface. In the next section, we describe how users can
 323 interact with the data resulting of those queries by means of an information visualization
 324 interface.

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

325 4.2 Visualization Interface

326 As mentioned earlier, LDViz uses the MGExplorer (**M**ultidimensional **G**raph
327 **E**xplorer) (Menin, Cava, et al., 2021) graphic library to support the visual exploration of
328 the Covid-on-the-Web dataset. More than a collection of charts, MGExplorer is a
329 visualization tool based on the concept of chained views, which supports the exploration of
330 multidimensional network data, while keeping provenance information to enable further
331 study of users' reasoning based on their interactions with the system. The visual
332 exploration process in MGExplorer consists of two phases, described as follows:

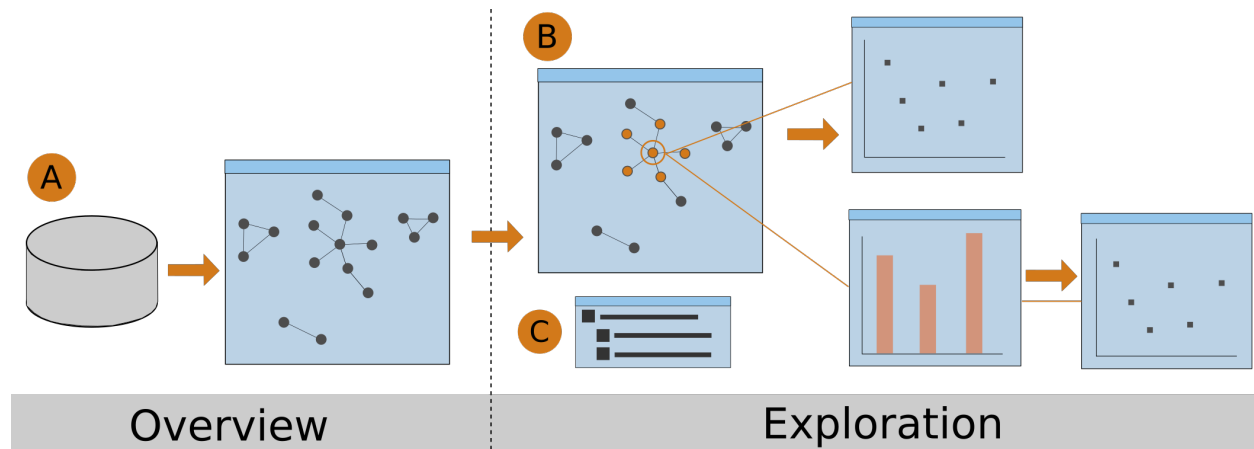
- 333 1. the *overview phase*, which consists of visualizing the network defined by the SPARQL
334 query results through a node-link diagram (see description below). This visualization
335 allows the user to get an overall understanding of the clusters within the data; and
- 336 2. the *exploratory phase*,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

339 The generic aspect of MGExplorer enables the combination of multiple
340 visualizations to support (1) the comparison of two or more different subsets of data
341 through a particular perspective provided by a particular visualization, and (2) the
342 comparison of different perspectives of the same subset of data using multiple,
343 complementary visualization techniques. Particularly, we currently support data
344 exploration through six views summarized in Table 1 and described as follows:

- 345 • The **node-link** diagram shows a set of nodes, which represent data items (e.g.,
346 authors), and their relationships represented through line segments connecting them.
347 In MGExplorer, this visualization technique provide an overview of the relationships
348 within items of the input data. In our use case scenarios (Section 5), the relationships
349 are defined by scientific publications, either to reveal co-authorship networks or
350 co-occurrence of named entities.

Figure 6

Overview of MGEplorer. (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).



- 351 • The **ClusterVis** technique (Cava et al., 2017) enables the inspection of clusters and

352 data attributes (e.g., publication type) within the subset of items (e.g., authors or

353 named entities). The visualization has a multi-ring layout, where the innermost ring

354 is formed by dots representing data items, and the remaining rings display the data

355 attributes, which can be customized and reordered by the user. The items in the

356 innermost ring that belong to the same sub-cluster are connected via curved lines,

357 which one can highlight by hovering over the items. The remaining rings are formed

358 by bars where height and color encode different data attributes (e.g., the height

359 encodes count and the color encodes the types of publications of a specific author).
- 360 • The **IRIS** technique represents the pairwise relationships between an item of interest

361 (e.g., an author) and the remaining items in a particular subset of data, which

362 relationship is described by data attributes (e.g., publication count and type) (Cava

363 et al., 2014). This technique is inspired by the eye's iris, which can only focus on a


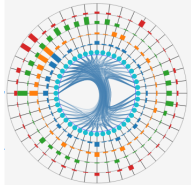
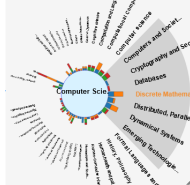
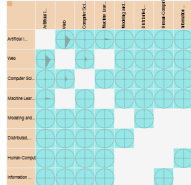
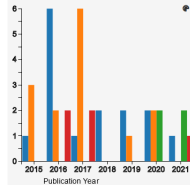

364 certain amount of information at the time, i.e., what is visible within our field of
365 view. The selected item is represented in the IRIS as a circle at the center of the
366 view, surrounded by its related items, which are displayed in a way that the ones in
367 the field of view (gray area) are larger than the ones outside this zone, easing
368 information extraction. The user can place any item in the field of view by clicking
369 on it, switching the focus of the IRIS. In order to represent data attributes describing
370 those pairwise relationships, we use the height and color of a bar placed in between
371 the item of interest and each of its related items.

- 372 • The **GlyphMatrix** technique (Cava & Freitas, 2013) features a matrix where rows
373 and columns represent data items (e.g., authors or named entities), and the
374 intersection cell between each pair of items contains a glyph encoding the data
375 attributes describing that relationship. The default glyph is based on a radar chart,
376 where each axis displays the count of a different data attribute (e.g., publication
377 type). The technique supports sorting of rows and columns to facilitate information
378 extraction, and hovering over cell to make the glyph larger and more visible through
379 a tooltip feature. This visualization technique could be seen as a combo of the
380 ClusterVis and IRIS by displaying the relationship between an item of interest and
381 other items in a pairwise manner, as well as the relationships within the remaining
382 items in the group.
- 383 • The **Bar chart** technique shows the distribution of publications according to a given
384 variable. In our case study, the x-axis encodes temporal information, while the y-axis
385 encodes the counting of publications. The data is displayed as a single bar per
386 time-period or multiple colored bars to represent categorical information of attributes.
- 387 • The **Listing** technique lists the items that form the relationship between two or more
388 nodes in the graph. In our case study, it displays the list of publications co-authored
389 by two or more authors or the publications where two or more named entities

390 co-occur, according to subset of data being explored. Each item of the list contains a
 391 link to a descriptive web page of the publication, where the user can obtain more
 392 information about it. Furthermore, if enabled by the GSS, each item contains a
 393 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
					
network	clusters	pairwise		distribution	listing

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

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

407 hiding currently displayed visualizations and/or showing any of the previous visualizations.

408 **5 Use Case Scenarios**

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

413 **5.1 Scenario 1. Clusters Visualization**

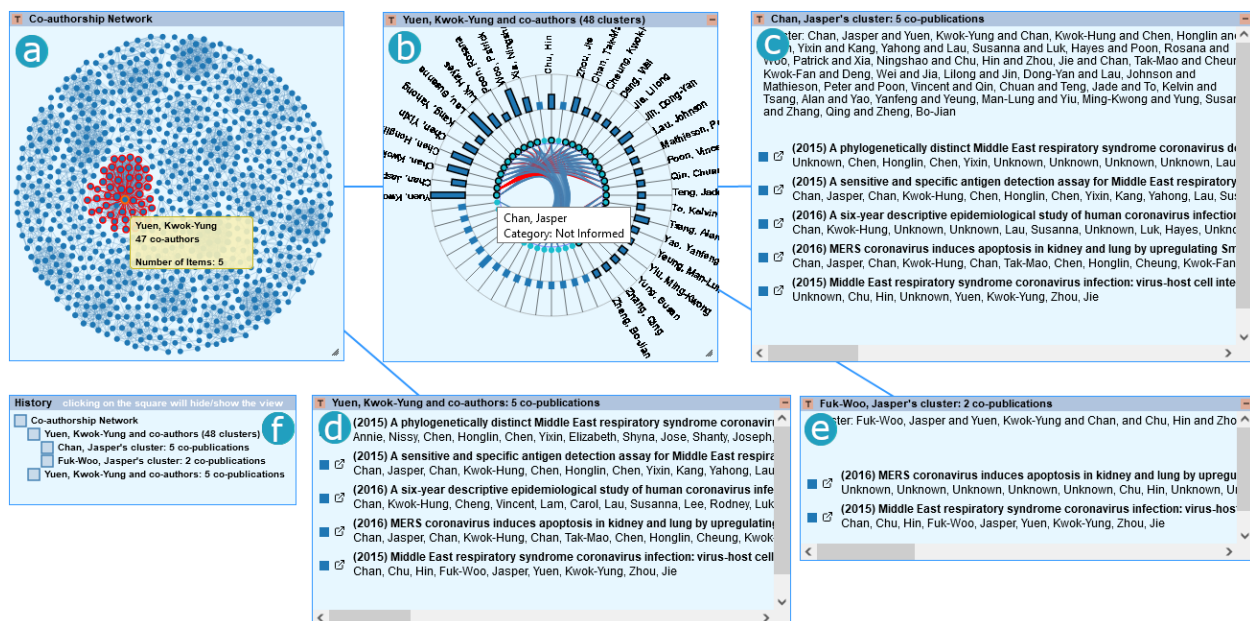
414 Based on the premise that COVID-19 has increased the collaboration between
415 researchers from diverse disciplines around the world (Naujokaitytė, 2021), a biomedical
416 researcher from INCa was interested on searching for information about existing
417 collaborations on the theme of the relationship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

421 We use a subset of data describing the co-authorship network within publications
422 related to coronavirus families retrieved with the query presented in Listing 1, which
423 resulted in 4,238 RDF triples corresponding to publications having the word “coronavirus”
424 in the title. These results were then transformed into a graph with 879 nodes (authors)
425 and 4,053 edges (connections between authors). Figure 7 depicts the exploratory path that
426 we follow during this scenario, which illustrates how one can explore clusters of co-authors
427 and related information to their co-publications. As mentioned earlier, the MGExplorer
428 visualization interface self-starts with an overview of co-authorship clusters through the
429 node-link diagram and the history tree of the exploratory process, which is progressively
430 completed based on the user’s interactions.

431 In the node-link diagram, we identify a dense sub-graph related to the author Yuen,
432 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.



433 over the node representing the author, where we observe that they have 47 co-authors,
 434 with whom five scholarly articles have been published. Subsequently, we right-click on the
 435 node to activate a context menu that allows subsetting the data and explore it with
 436 another visualization technique. We choose the ClusterViz view, where we can explore the
 437 different clusters within the subset of co-authors selected in the node-link (Figure 7c). For
 438 two different clusters, we subset the data by hovering over a particular author and display
 439 the list of publications which they co-authored together (Figure 7d-e). Finally, we could
 440 compare the contributions made within those clusters and the complete list of publications
 441 co-authored by our author of interest (Figure 7f), to understand the impact of these
 442 co-authorship relationships in terms of number and quality of publications they have

443 together.

```

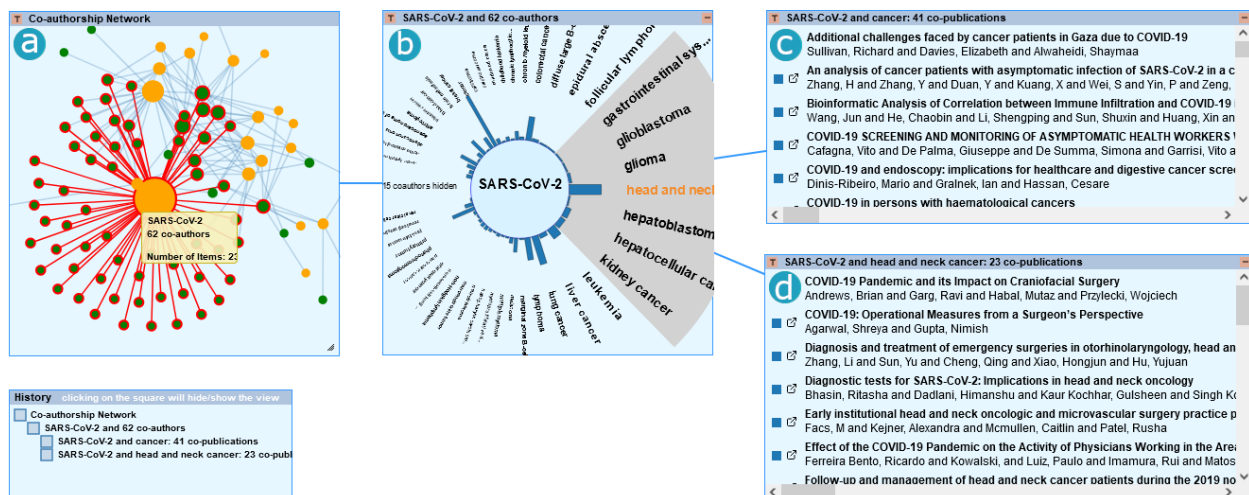
444 select ?p (group_concat(distinct ?name ; separator = '--') as ?author)
445     (sample(?d) as ?date) (sample(?t) as ?label) where {
446     select * where {
447         ?doc dct:title ?t ; dct:issued ?d ; dce:creator ?name .
448         filter contains(?t, "coronavirus")
449         filter (?d >= "$beginYear-01-01"^^xsd:date) # $beginYear = 2015
450         filter (?d <= "$endYear-12-31"^^xsd:date) # $endYear = 2021
451     } limit 1000
452 } 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.

453 **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.



454 The generic structure of LDViz allows the construction of graphs with different
 455 topologies. The user can choose the variables that correspond to the nodes and the
 456 connection between them (e.g., in the previous scenario, nodes correspond to a variable
 457 that describes the authors' names and the edges correspond to a variable that describe the
 458 documents they co-authored). Together with biomedical researchers, we have identified the
 459 task “to identify the articles that mention both a type of cancer and a virus of the corona
 460 family” as being relevant for their analyses. Thus, in this scenario, we illustrate how we
 461 can use the LDViz to solve this domain-related task. Using the query presented in
 462 Listing 3, we retrieve the RDF triples that correspond to the pattern $?s \rightarrow ?p \rightarrow ?o$, where
 463 $?s$ and $?o$ are, respectively, named entities related to (i.e., equal to, subclass of, or instance
 464 of) “cancer” and “coronavirus” named entities, and $?p$ refers to the publications that
 465 contain these named entities on their text body. The relationships are determined by
 466 publications, however, unlike the Scenario 1, this query modifies the topology of the graph
 467 to represent the relationships between named entities instead of co-authors.

```
468 {"node": { "fst": {"color": "green"}, "snd": {"color": "orange"} },
469 "services": [{"label": "ACTA", "url": "http://134.59.134.234:8081/analyseddocs?search="},
470 {"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

```
471 # wdt:P279 = subclass of, wdt:P31 = instance of
472 # wd:Q1134583 = coronavirus family, wd:Q12078 = cancer
473 prefix wd: <http://www.wikidata.org/entity/>
474 prefix wdt: <http://www.wikidata.org/prop/direct/>
475
476 select distinct ?s ?p ?o ?label ?pmid ?authorList ("fst" as ?style1) ("snd" as ?style2)
477 from <http://ns.inria.fr/covid19/graph/entityfishing>
478 from <http://ns.inria.fr/covid19/graph/articles>
479 from named <http://ns.inria.fr/covid19/graph/wikidata-named-entities-full>
480 where {
481     ?annot1 schema:about ?p ; oa:hasBody ?dis1.
482     ?annot2 schema:about ?p ; oa:hasBody ?dis2.
483     ?p dct:title ?label ; bibo:pmid ?pmid .
484     graph <http://ns.inria.fr/covid19/graph/wikidata-named-entities-full> {
```

```

485     {?dis1 rdfs:label ?s. filter(?dis1=wd:Q12078)} UNION
486     {?dis1 wdt:P279 wd:Q12078; rdfs:label ?s.} UNION {?dis1 wdt:P31 wd:Q12078; rdfs:label ?s.}
487     {?dis2 rdfs:label ?o. filter(?dis2=wd:Q1134583)}
488     UNION {?dis2 wdt:P279 wd:Q1134583; rdfs:label ?o.} }
489     {select ?p (group_concat(?name ; separator="--") as ?authorList) where {
490         ?p dce:creator ?name
491     } group by ?p}
492 } 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.

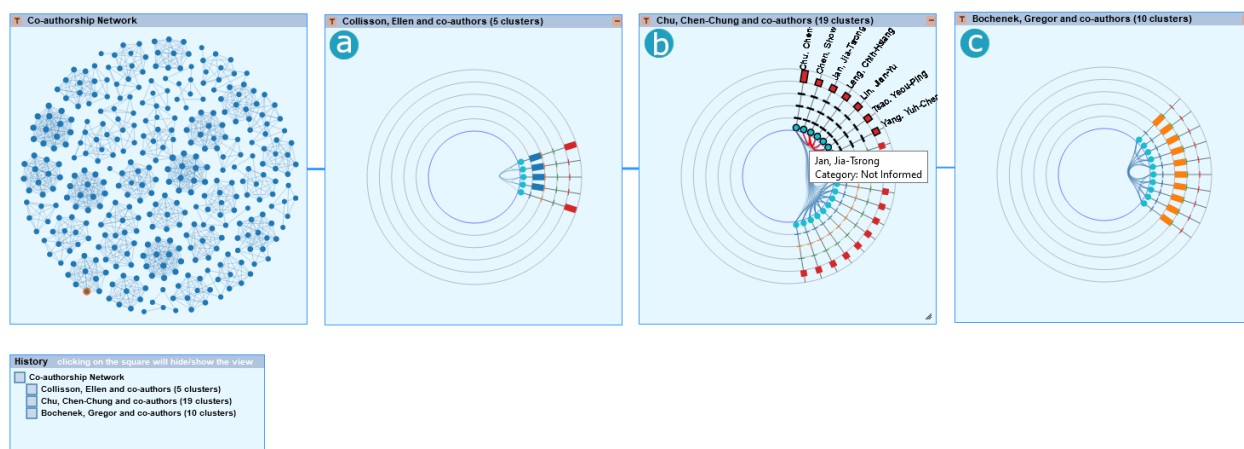
493 Figure 8 depicts the exploratory path followed in this scenario to solve the
494 above-described domain-related task. We explore a dataset that contains 452 RDF triples,
495 which results in a graph with 94 nodes and 169 edges. Since in this dataset, we deal with
496 two types of nodes, i.e., either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

511 information about each document²².

512 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).



513 The previous exploration scenarios allow the user to see the relationship between
 514 co-authors or named entities, which can be characterized by the number of related
 515 publications. Thus, this scenario illustrates how we can use the LDViz to explore custom
 516 data attributes of a co-authorship network within coronavirus-related publications. In
 517 particular, we will use a dataset that describes publications through the research topic
 518 retrieved with the query presented in Listing 4. In the context of the Covid-on-the-Web
 519 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>

<http://ns.inria.fr/covid19/28ecacb70247f4fb6a4923a99d0905153c23f88a>

520 of named entities that can be used to describe the research topic of the publication. The
 521 resulting dataset has 1,265 RDF triples, which were transformed in a graph with 356 nodes
 522 (authors) and 1,262 edges (co-publications). From the resulting data, the system identified
 523 the values “sequence alignment”, “reverse transcriptase”, and “transfection” as the most
 524 relevant research topics to describe the publications within the data and classified the
 525 remaining under the “other” category.

526 Figure 9 depicts the exploratory path of this scenario. We inspect the clusters of
 527 co-authorship within the associations of different authors through the ClusterViz
 528 visualization. We can observe, for instance, that the researcher Collisson, Ellen (Figure 9a)
 529 has publications about different topics (i.e., sequence alignment and other) within different
 530 clusters of co-authorship, while the publications co-authored by Chu, Chen-Chung
 531 (Figure 9b) refer to the “other” category of topics and are distributed throughout different
 532 clusters of co-authorship. Finally, we observe that the publication co-authored by
 533 Bochenek, Gregor (Figure 9c), for instance, refers to the topic of “reverse transcriptase”.

```

534 select ?p (group_concat(distinct ?name ; separator = '--') as ?author)
535     (sample(?d) as ?date) (sample(?t) as ?label)
536     (sample(?label) as ?type) where {
537     select * where {
538         ?p dct:title ?t ; dct:issued ?d ; dce:creator ?name .
539         filter contains(?t, "coronavirus")
540
541         graph <http://ns.inria.fr/covid19/graph/entityfishing> {
542             ?a1 a oa:Annotation; schema:about ?p ; oa:hasBody ?uri . }
543             ?uri rdfs:label ?subject .
544             FILTER (langMatches( lang(?subject), "EN" ) )
545         } limit 10000
546     } 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.

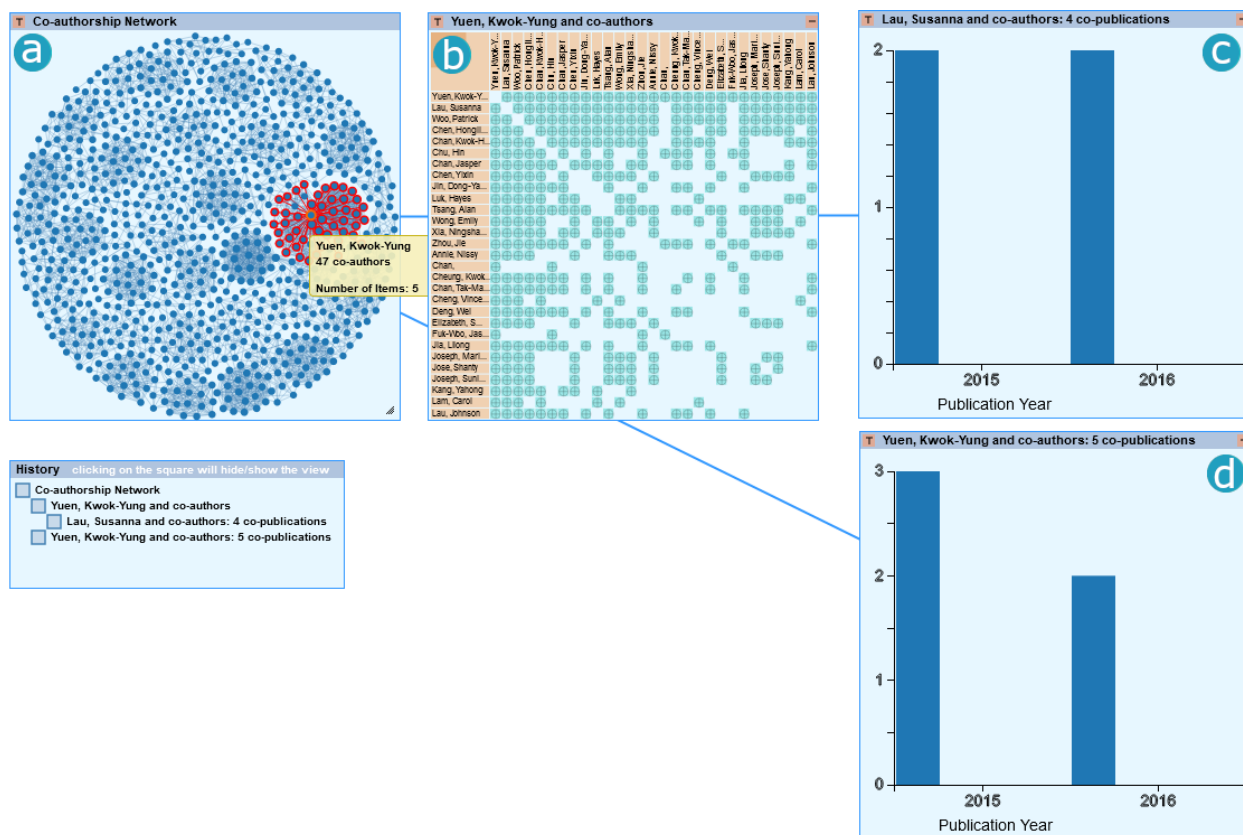
547 **5.4 Scenario 4. Exploring the Temporal Aspect of Relationships**

548 Studying the evolution over time of co-authors relationships or named entities
549 co-occurrence could help understand when a collaboration between authors were stronger
550 or when certain research topics were of higher interest, which information could be further
551 explained with context, e.g., nowadays the research around the coronavirus topic is
552 stronger than ever due to the COVID-19 pandemics. Thus, in this scenario, we illustrate
553 how one can use the LDViz interface to explore the temporal aspects of relationships,
554 particularly co-authorship within publications related to coronaviruses (see Listing 1).

555 Figure 10 depicts the exploratory path used in this scenario. Similar to Scenario 1,
556 we use the node-link diagram to identify the author with the most co-authors, i.e., Yuen,
557 Kwok-Yung (hereafter called author A) with 47 co-authors associated through five
558 publications (Figure 10a). We further explore the relationship between author A and their
559 co-authors through the GlyphMatrix visualization, which shows the types and number of
560 co-publications between author A and every other co-author, as well as the co-publications
561 among author A's co-authors. By ordering rows and columns by the number of
562 co-publications, we can observe in the GlyphMatrix, that author A's most recurrent
563 co-author is Lau, Susanna (hereafter called author B) (Figure 10b), with whom they have 4
564 publications. Thus, to verify when these collaborations happened, we explore the temporal
565 distribution of co-publications between those authors by subsetting the data in the
566 GlyphMatrix visualization and exploring it on the Histogram technique (Figure 10c). We
567 observe that they had collaborations in 2015 and 2016. When comparing to the totality of
568 co-publications related to author A (Figure 10d), we observe that four out of five
569 publications are co-authored by author B which could indicate a strong collaboration
570 between those authors in co-publications related to the coronavirus topic. We can also
571 observe that this collaboration appear to have ended five years ago, since the dataset
572 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.

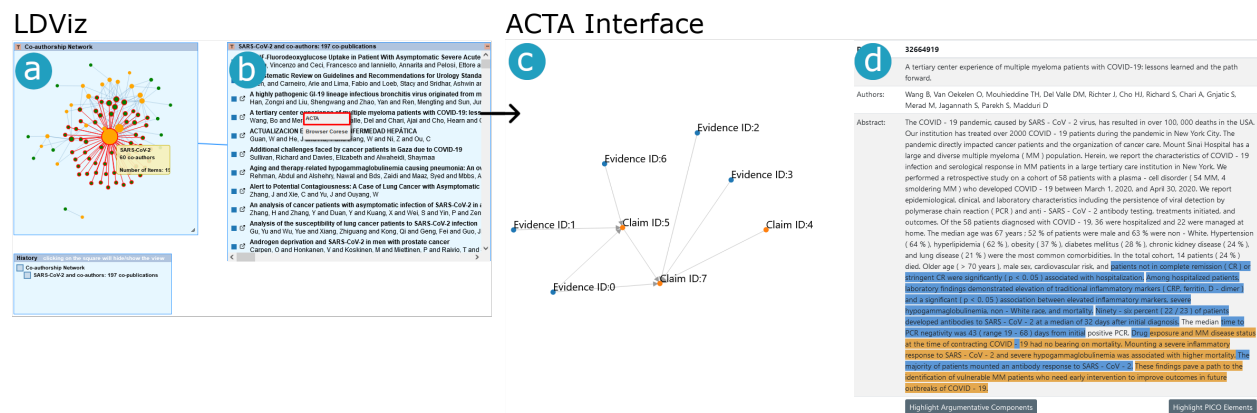


573 **5.5 Scenario 5. Exploring Argumentation Graphs with the ACTA Interface**

574 As mentioned earlier, the GSS feature allows the user to include external services in
 575 LDViz, such as a service that enable a further exploration of the resources currently being
 576 visualized with the LDViz interface. In this scenario, we explore the subset of data used in
 577 Scenario 2 (i.e., set of publications where named entities related to “cancer” and
 578 “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.



579 argumentative graph of a certain publication identified during the exploratory process. As
 580 one can see in Listing 2, the GSS form associate to the query contains an object called
 581 “services” that provides the redirection information for the ACTA interface (i.e., a call to
 582 “http://134.59.134.234:8081/analyseddocs?search=”). The documents used in the
 583 Covid-on-the-Web dataset often originate from the PubMed archive²³, where each
 584 document has a unique identifier. Thus, upon the selection of a document, the LDViz
 585 system launches the ACTA service by redirecting the user to the given URL, while
 586 providing the document identifier as parameter.

587 Figure 11 depicts the exploratory path used in Scenario 5. As for Scenario 2, we
 588 identify the larger sub-graph in the node-link diagram, which is the one connecting to the
 589 node that corresponds to the named entity “SARS-Cov-2” (Figure 11a). Using the

²³ <https://pubmed.ncbi.nlm.nih.gov/>

590 Histogram, we display the 232 publications where this named entity occurs (Figure 11b).
591 Subsequently, we can choose any of the listed publications for which we would like to
592 visualize the argumentative graph using ACTA. We right-click on the publication of
593 interest and choose the “ACTA” option on the context menu that appears. This action
594 redirects the user to the ACTA interface, which retrieves the selected document from the
595 PubMed server, analyzes it, and display the resulting argumentative graph with the
596 relationships between claims and evidences, and PICO elements (Figure 11c). One can also
597 inspect these elements using the textual information (Figure 11d), where we can choose to
598 highlight the argumentative sentences or the PICO elements. Alternatively, one can query
599 the CORD19-AKG²⁴ dataset to explore claims and evidences graph related to one or more
600 publications directly on LDViz by using a SPARQL query where `?s` and `?o` correspond to
601 claims and evidences, while the `?p` variable correspond to the publication(s) where they
602 were identified.

603 **6 Discussion**

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

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

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

615 the Covid-on-the-Web dataset from different perspectives, enabling the resolution of
616 various domain-related questions. In this paper, we have particularly focused on the LDViz
617 visualization tool, which supports the visual exploration of subsets of data defined by
618 SPARQL queries. The tool is based on the MGExplorer visualization framework, which
619 proposes a collection of charts linked together through a chained visualization approach
620 that allows us to keep track of the exploration path, assisting the understanding of the
621 sensemaking process. This visualization aims to help users understand the relationships
622 within the results, e.g., users can run a query to visualize a co-authorship network; then
623 use IRIS and ClusterVis to understand who is working together and on which research
624 topics. An interesting aspect of our approach is that one can change the graph topology to
625 explore relationships between different kinds of items. For instance, the user could execute
626 a query that looks for papers mentioning the COVID-19 and diverse types of cancer, as
627 illustrated in the Use Case Scenario 2 (see Subsection 5.2). Another strong aspect of
628 LDViz relies on the possibility of exploring the relationships within any subset of data
629 originating from any SPARQL endpoint thanks to the data transformation engine that
630 adapts the query's results to the data format required by the visualization.

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

640 Although a first level of evaluation is shown by translating the user queries to
641 SPARQL queries to visual data in LDViz, which shows that our dataset and visualization

642 services support the resolution of users queries, user evaluations are essential to validate
643 the usability and utility of a visualization. However, evaluating LDViz (as well as any
644 visualization) is not a trivial task since it has been designed to support exploratory tasks,
645 which are the hardest ones to replicate in an experiment (Ellis & Dix, 2006). Furthermore,
646 the value of LDViz can only be assessed when used by professionals on the application
647 domain (e.g., biomedical researchers), who are difficult to recruit since they are not
648 necessarily available to take part in experiments. Future work includes implementing
649 user-based evaluations to investigate the usability of LDViz tool for exploring linked
650 datasets in general, and in particular its suitability for analyzing the COVID-19 scientific
651 literature and assisting the resolution of domain-related tasks.

652 The generic aspects of our tools allow us to later on apply the resources to a wider
653 set of use case scenarios, which possibility have been evoked by our biomedical partners,
654 who would like to perform similar analyses over issues other than the COVID-19. In fact,
655 the LDViz interface has been applied to two other publication datasets (i.e., HAL open
656 archive²⁵ and the Microsoft Academic Knowledge Graph²⁶, for which a set of predefined
657 queries are available at <http://covid19.i3s.unice.fr:8080/hal>). The genericity of our
658 approach enables the exploration of data from any SPARQL endpoint, such as DBpedia²⁷,
659 from which we explored the ontology and RDF Schema information, as well as a
660 co-starring relationship using movies information²⁸. The tool also has a generic service that
661 enables the querying and visualization of any SPARQL endpoint, which URL can embed a
662 SPARQL query and the URL of a SPARQL endpoint²⁹, to directly visualize the resulting
663 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>>

664 service³⁰ to combine data from different SPARQL endpoints using federated queries.

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

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

³⁰ <http://corese.inria.fr/sparql>

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

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

712 **7 Conclusion**

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

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

715 support the exploration of the COVID-19 scientific literature. Based on the needs of
716 biomedical researchers, partners of the project, we designed and published a linked data
717 knowledge graph describing the named entities mentioned in the articles of the COVID-19
718 corpus and the argumentative graphs they include. The knowledge graph generation
719 pipeline has been published to allow the scientific community to reuse, enrich and adapt
720 both the dataset and the pipeline in meaningful ways to assist users needs.

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

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

734 **Acknowledgments**

735 This work is partly funded by the French government labelled PIA program under
736 its IDEX UCAJEDI project (ANR-15-IDEX-0001) and the 3IA Côte d'Azur
737 (19-P3IA-0002) as well as the CovidOnTheWeb project funded by Inria. We gratefully
738 acknowledge the contributions of Valentin Ah-Kane and Mathieu Simon and our research
739 partners of Inserm and INCa institutions. We also acknowledge the contribution of Carla
740 Freitas and Ricardo Cava for the initial work on the MGExplorer framework.

741 **References**

- 742 Ambavi, H., Vaishnav, K., Vyas, U., Tiwari, A., & Singh, M. (2020). Covidexplorer: A
743 multi-faceted ai-based search and visualization engine for covid-19 information.
744 *Proceedings of the 29th ACM International Conference on Information & Knowledge*
745 *Management*, 3365–3368. <https://doi.org/10.1145/3340531.3417428>
- 746 Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: Pretrained Language Model for
747 Scientific Text. *EMNLP, arXiv preprint arXiv:1903.10676*.
- 748 Bras, P. L., Gharavi, A., Robb, D. A., Vidal, A. F., Padilla, S., & Chantler, M. J. (2020).
749 Visualising covid-19 research. *arXiv preprint arXiv:2005.06380*.
- 750 Cava, R., & Freitas, C. D. S. (2013). Glyphs in matrix representation of graphs for
751 displaying soccer games results. *The 1st Workshop on Sports Data Visualization.*
752 *IEEE*, 13, 15. <http://workshop.sportvis.com/papers/cavaSoccerMatches.pdf>
- 753 Cava, R., Freitas, C. M. D. S., & Winckler, M. (2017). Clustervis: Visualizing nodes
754 attributes in multivariate graphs. *Proceedings of the Symposium on Applied*
755 *Computing*, 174–179. <https://doi.org/10.1145/3019612.3019684>
- 756 Cava, R., Freitas, C. M., Barboni, E., Palanque, P., & Winckler, M. (2014). Inside-in
757 search: An alternative for performing ancillary search tasks on the web. *2014 9th*
758 *Latin American Web Congress*, 91–99. <https://doi.org/10.1109/LAWeb.2014.21>
- 759 Corby, O., Gaignard, A., Faron-Zucker, C., & Montagnat, J. (2012). KGRAM Versatile
760 Data Graphs Querying and Inference Engine. *Proc. IEEE/WIC/ACM International*
761 *Conference on Web Intelligence*. <https://dl.acm.org/doi/10.5555/2457524.2457672>
- 762 Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and
763 accuracy in multilingual entity extraction. *Proceedings of the 9th International*
764 *Conference on Semantic Systems*, 121–124.
765 <https://doi.org/10.1145/2506182.2506198>
- 766 Ellis, G., & Dix, A. (2006). An Explorative Analysis of User Evaluation Studies in
767 Information Visualisation. *Proceedings of the 2006 AVI Workshop on BEyond Time*

- 768 *and Errors: Novel Evaluation Methods for Information Visualization*, 1–7.
769 <https://doi.org/10.1145/1168149.1168152>
- 770 Fonseca, B. d. P. F. e., Sampaio, R. B., de Araújo Fonseca, M. V., & Zicker, F. (2016).
771 Co-authorship network analysis in health research: Method and potential use.
772 *Health Research Policy and Systems*, 14(1), 1–10. [https://health-policy-](https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-016-0104-5)
773 [systems.biomedcentral.com/articles/10.1186/s12961-016-0104-5](https://health-policy-systems.biomedcentral.com/articles/10.1186/s12961-016-0104-5)
- 774 Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., Hearst, M. A., &
775 West, J. (2020). SciSight: Combining faceted navigation and research group
776 detection for COVID-19 exploratory scientific search. *arXiv preprint*
777 *arXiv:2005.12668*.
- 778 Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N. T., Yao, Y., Rogers, C., Li, R., Liu, J.,
779 Singh, A., Schwabe, D., & Szekely, P. (2020). KGTK: A Toolkit for Large
780 Knowledge Graph Manipulation and Analysis. *The Semantic Web – ISWC 2020*,
781 278–293. https://doi.org/10.1007/978-3-030-62466-8_18
- 782 Jonquet, C., Shah, N. H., & Musen, M. A. (2009). The open biomedical annotator. *Summit*
783 *on translational bioinformatics, 2009*, 56.
784 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041576/>
- 785 Leng, J. (2011). *Handbook of research on computational science and engineering: Theory*
786 *and practice* (Vol. 2). IGI Global.
- 787 Lohmann, S., Negru, S., Haag, F., & Ertl, T. (2016). Visualizing ontologies with VOWL.
788 *Semantic Web*, 7(4), 399–419. <https://doi.org/10.3233/SW-150200>
- 789 Mayer, T., Cabrio, E., & Villata, S. (2019). ACTA a tool for argumentative clinical trial
790 analysis. *Proceedings of the 28th International Joint Conference on Artificial*
791 *Intelligence (IJCAI)*, 6551–6553. <https://doi.org/10.24963/ijcai.2019/953>
- 792 Menin, A., Cadorel, L., Tettamanzi, A., Giboin, A., Gandon, F., & Winckler, M. (2021).
793 ARViz: Interactive Visualization of Association Rules for RDF Data Exploration.
794 *25th International Conference Information Visualisation*.

- 795 Menin, A., Cava, R., Freitas, C. M. D. S., Corby, O., & Winckler, M. (2021). Towards a
796 Visual Approach for Representing Analytical Provenance in Exploration Processes.
797 *25th International Conference Information Visualisation*.
- 798 Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., Corby, O., Gazzotti, R.,
799 Giboin, A., Marro, S., Mayer, T., Simon, M., Villata, S., & Winckler, M. (2020).
800 Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research.
801 In J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres,
802 O. Seneviratne, & L. Kagal (Eds.), *The Semantic Web – ISWC 2020* (pp. 294–310).
803 Springer International Publishing. https://doi.org/10.1007/978-3-030-62466-8_19
- 804 Naujokaitytė, G. (2021). COVID-19 triggered unprecedented collaboration in research
805 [Accessed on 6 July 2021].
- 806 Oniani, D., Jiang, G., Liu, H., & Shen, F. (2020). Constructing co-occurrence network
807 embeddings to assist association extraction for covid-19 and other coronavirus
808 infectious diseases. *Journal of the American Medical Informatics Association*, *27*(8),
809 1259–1267. <https://doi.org/10.1093/jamia/ocaa117>
- 810 Pietriga, E. (2006). Semantic web data visualization with graph style sheets. *Proceedings of*
811 *the 2006 ACM symposium on Software visualization*, 177–178.
812 <https://doi.org/10.1145/1148493.1148532>
- 813 Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S.,
814 Shefchek, K. A., Good, B. M., Balhoff, J. P., Fontana, T., et al. (2021).
815 KG-COVID-19: a framework to produce customized knowledge graphs for
816 COVID-19 response. *Patterns*, *2*(1), 100155.
817 <https://doi.org/10.1016/j.patter.2020.100155>
- 818 Sukla, A., Naskar, A., Goel, T., Sangwan, S., Rai, A., Shakir, M., Verma, I., Dasgupta, T.,
819 & Dey, L. (2021). Concept Driven Search and Visualization System for Exploring
820 Scientific Repositories. *8th acm ikdd cods and 26th comad* (pp. 395–399).
821 <https://doi.org/10.1145/3430984.3430991>

- 822 Tu, J., Verhagen, M., Cochran, B., & Pustejovsky, J. (2020). Exploration and discovery of
823 the COVID-19 literature through semantic visualization. *arXiv preprint*
824 *arXiv:2007.01800*.
- 825 Verspoor, K., Šuster, S., Otmakhova, Y., Mendis, S., Zhai, Z., Fang, B., Lau, J. H.,
826 Baldwin, T., Yepes, A. J., & Martinez, D. (2020). COVID-SEE: Scientific Evidence
827 Explorer for COVID-19 related research. *arXiv preprint arXiv:2008.07880*.
- 828 Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K.,
829 Kinney, R. M., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D.,
830 Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wilhelm, C., ...
831 Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. *ArXiv*,
832 *abs/2004.10706*.

833 **8 Author contributions (Contributor Roles Taxonomy Project – CRediT)**

834 **Aline Menin:** Conceptualization, Investigation, Methodology, Software, Writing –
835 original draft, Writing – review & editing

836 **Franck Michel:** Data Curation, Investigation, Resources, Software, Writing – review &
837 editing

838 **Fabien Gandon:** Funding Acquisition, Writing – review & editing

839 **Raphael Gazzotti:** Resources, Writing – review & editing

840 **Elena Cabrio:** Supervision

841 **Olivier Corby:** Software

842 **Alain Giboin:** Investigation, Methodology, Writing – review & editing

843 **Santiago Marro:** Resources

844 **Tobias Mayer:** Resources

845 **Serena Villata:** Supervision

846 **Marco Winckler:** Conceptualization, Writing – original draft, Writing – review &
847 editing, Supervision, Methodology, Formal analysis

848

849

The authors have no competing interests regarding this work.