Covid-on-the-Web: Graphe de Connaissances et Services pour faire Progresser la Recherche sur la COVID-19

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

Covid-on-the-Web: Graphe de Connaissances et Services pour faire Progresser la Recherche sur la COVID-19

F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby,
R. Gazzotti, A. Giboin, S. Marro, T. Mayer, M. Simon, S. Villata, M. Wincker.

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

franck.michel@cnrs.fr

Résumé

Le projet Covid-on-the-Web permet aux chercheurs d'accéder à la littérature relative à la famille des coronavirus, de l'interroger et d'en extraire des connaissances. Il s'aligne sur des besoins concrets formulés par des instituts de santé et de recherche. Ainsi, il adapte, combine et étend des outils destinés à traiter, analyser et enrichir le corpus CORD-19 qui rassemble plus de 100 000 articles scientifiques relatifs aux coronavirus. Ce jeu de données comprend deux principaux graphes de connaissances décrivant (1) 113 millions de mentions d'entités nommées liées au Web de données, et (2) les arguments extraits à l'aide d'ACTA, un outil d'extraction et de visualisation de graphes argumentatifs. Nous fournissons également plusieurs outils de visualisation et d'exploration basés sur la plateforme Corese, la bibliothèque MGExplorer, ainsi que des Notebooks Jupyter.

Mots-clés

COVID-19, arguments, visualisation, entités nommées, données liées.

Abstract

The Covid-on-the-Web project allows scientists to access, query and extract knowledge from the literature on the coronavirus family. It is aligned with concrete needs formulated by health and research institutes. Thus, it adapts, combines and extends tools designed to process, analyze and enrich the CORD-19 corpus, that gathers 100,000+ scientific articles related to the coronaviruses. This dataset comprises two main knowledge graphs describing (1) 113 million mentions of named entities linked to the Web of data, and (2) arguments extracted using ACTA, a tool for extraction and visualization of argumentative graphs. We also provide several visualization and exploration tools based on the Corese platform, the MGExplorer library, and Jupyter Notebooks.

Keywords

COVID-19, arguments, visualization, named entities, linkeddata.

1 Des données sur la COVID-19 vers des données ouvertes liées

En mars 2020, alors que la maladie infectieuse respiratoire COVID-19 nous obligeait à rester confinés, l'équipe de recherche Wimmics [1] a décidé de se joindre aux efforts de nombreux scientifiques du monde entier qui mettent à profit leur expertise et leurs ressources pour lutter contre la pandémie et en atténuer ses effets dévastateurs. Nous avons lancé un nouveau projet nommé *Covid-on-the-Web* visant à faciliter l'accès, la recherche et la compréhension de la littérature scientifique biomédicale relative au COVID. À cette fin, nous avons adapté, réorienté, combiné et utilisé des outils pour publier, aussi exhaustivement et rapidement que possible, un maximum de données liées relatives aux coronavirus.

En quelques semaines, nous avons déployé plusieurs outils afin d'analyser le *COVID-19 Open Research Dataset* (CORD-19) [18] qui compte plus de 100 000 articles scientifiques relatifs à la famille des coronavirus. D'une part, nous avons adapté la plateforme ACTA, [2] conçue initialement pour aider les cliniciens dans l'analyse des essais cliniques et la prise de décision [11], en permettant l'extraction automatique et la visualisation des graphes argumentatifs. D'autre part, notre expertise dans la gestion des données extraites à l'aide de graphes de connaissances, qu'elles soient génériques ou spécialisées, et leur intégration dans le projet HealthPredict [8, 9], nous ont permis d'enrichir le corpus CORD-19 avec différentes sources. Nous avons utilisé DBpedia Spotlight [5], Entity-fishing [3] et NCBO BioPortal Annotator [10] afin d'extraire les entités nommées des articles du corpus CORD-19, et les désambiguïser en regard des ressources des données ouvertes liées venant de DBpedia, Wikidata et BioPortal. En utilisant la plateforme Morph-xR2RML, [4] nous avons transformé le résultat en un jeu de données RDF que nous avons publié via un point d'accès SPARQL public. En parallèle, nous avons intégré les plateformes Corese [5] [4] et MGExplorer [3] pour mani-

1. https://team.inria.fr/wimmics/
2. http://ns.inria.fr/acta/
3. https://github.com/kermitt2/entity-fishing
4. https://github.com/frmichel/morph-xr2rml/
5. https://project.inria.fr/corese/

puler des graphes de connaissances et permettre leur visualisation et leur exploration sur le web.

Le projet Covid-on-the-Web (représenté dans la Figure 1) a ainsi conçu et mis en place un pipeline (workflow) intégré facilitant l'extraction et la visualisation des informations issues du corpus CORD-19 par la production et la publication d'un graphe de connaissances de données liées enrichi en permanence. Nous pensons que notre approche, qui intègre des structures argumentatives et des entités nommées, est pertinente dans le contexte actuel. En effet, alors que de nouvelles recherches liées à la COVID-19 sont publiées chaque jour, les résultats sont activement débattus, et de nombreuses controverses voient le jour (sur l'origine de la maladie, son diagnostic, son traitement...) [2]. Les chercheurs ont donc besoin d'outils pour les aider à étayer ou écarter certaines hypothèses, traitements ou explications. L'exploitation conjointe de structures argumentatives et de raisonnement basé sur les entités nommées peut aider à répondre aux besoins de ces utilisateurs et ainsi réduire les zones d'ombres liées à la maladie.

Cet article est un résumé long traduit et mis à jour de l'article [14] que nous avons publié à ISWC 2020 et dans lequel nous dressons un bilan ainsi qu'une comparaison avec les travaux connexes (non reproduits dans cet article).

Le reste de cet article est organisé de la manière suivante. Dans la section 2, nous détaillons le pipeline d'extraction mis en place pour traiter le corpus CORD-19 et générer des données RDF. Puis, la section 3 détaille les caractéristiques du jeu de données et des services mis à disposition pour l'exploiter. Les sections 4 et 5 présentent les outils d'exploitation et de visualisation, et traitent des applications futures et de l'impact potentiel de notre jeu de données.

2 De CORD-19 au jeu de données Covid-on-the-Web

Le *COVID-19 Open Research Dataset* [18] (CORD-19) est un corpus rassemblant des articles scientifiques liés au SARS-Cov-2 et à la famille des coronavirus. Les créateurs de CORD-19 ont traité plus de 100 000 articles et les ont convertis en documents JSON tout en nettoyant les citations et les références bibliographiques.

Cette section décrit comment nous avons exploité ce jeu de données pour *(1)* établir des liens significatifs entre les articles du corpus CORD-19 et le Web de données au moyen des entités nommées, et *(2)* extraire un graphe d'arguments découverts dans les articles, tout en reposant sur les normes du Web sémantique et les pratiques des données liées.

2.1 Construction du graphe de connaissances des entités nommées CORD-19

Le graphe de connaissances des entités nommées CORD-19 (CORD19-NEKG), décrit les entités nommées identifiées et désambiguïsées dans les articles du corpus CORD-19 à l'aide de trois logiciels : DBpedia Spotlight [5] pour désambiguïser et lier les ENs à DBpedia dont nous avons utilsé les modèles pré-entraînés [6] ; Entity-fishing [7] pour désambiguïser les ENs en regard à Wikidata ; et NCBO BioPortal Annotator [8] [10] qui permet d'annoter du texte biomédical et désambiguïser les ENs en regard des ontologies se trouvant sur BioPortal.

Listing 1 – Représentation de l'entité nommée "réaction en chaîne par polymérase" (PCR) comme une annotation du résumé d'un article présente de la position 235 à 238.

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

[] a oa:Annotation;
  schema:about <http://ns.inria.fr/covid19/f74923b3...>;
  dct:subject "Engineering", "Biology";
  covidpr:confidence   "1"^^xsd:decimal;
  oa:hasBody <http://wikidata.org/entity/Q176996>;
  oa:hasTarget [
    oa:hasSource
    <http://ns.inria.fr/covid19/f74923b3...#abstract>;
    oa:hasSelector [
      a oa:TextPositionSelector, oa:TextQuoteSelector;
      oa:exact "PCR"; oa:start "235"; oa:end "238" ]];
```

Pour assurer sa réutilisabilité, CORD19-NEKG s'appuie sur des vocabulaires largement répandus, adaptés à la représentation des articles et des entités nommées en RDF. Nous présentons ci-dessous les principaux concepts de cette modélisation. De plus amples détails sont disponibles sur le dépôt Github du projet. [9]

Les métadonnées (ex. titre, auteurs, DOI) et le contenu des articles sont décrits à l'aide de DCMI, [10] FRBR-aligned Bibliographic Ontology (FaBiO), [11] Bibliographic Ontology, [12] FOAF [13] et Schema.org. [14] Les entités nommées sont représentées comme des annotations à l'aide du Web Annotation Vocabulary. [15] Un exemple d'annotation est donné dans le Listing 1. Le corps de l'annotation (`oa:hasBody`) correspond à l'URI de la ressource liée à l'entité détectée. Le morceau de texte reconnu comme l'entité nommée est la cible de l'annotation (`oa:hasTarget`). Celle-ci indique la partie de l'article dans laquelle l'entité a été reconnue (titre, résumé ou corps de l'article), et en donne sa position. Chaque annotation est accompagnée d'informations de provenance exprimées à l'aide de l'ontologie PROV-O, [16] indiquant la source en cours de traitement, l'outil utilisé pour extraire l'entité, le degré de confiance de l'annotateur sémantique, ainsi que l'auteur de l'annotation.

6. https://downloads.dbpedia.org/repo/dbpedia/spotlight/spotlight-model/
7. https://github.com/kermitt2/entity-fishing
8. http://data.bioontology.org/documentation
9. https://github.com/Wimmics/covidontheweb
10. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
11. https://sparontologies.github.io/fabio/current/fabio.html
12. http://bibliontology.com/specification.html
13. http://xmlns.com/foaf/spec/
14. https://schema.org/
15. https://www.w3.org/TR/annotation-vocab/
16. https://www.w3.org/TR/prov-o/

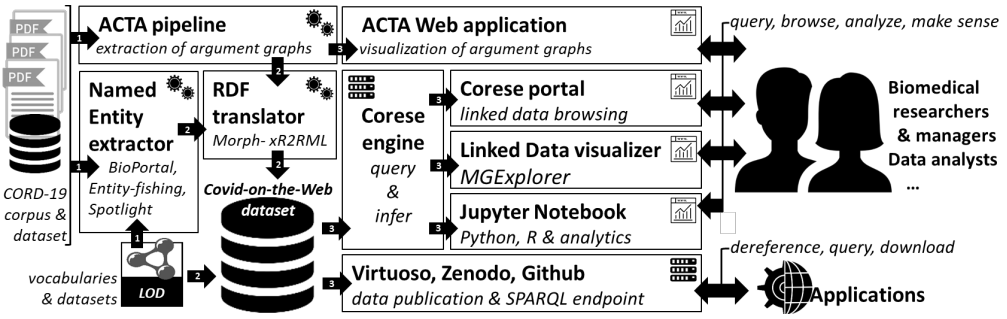

FIGURE 1 – Vue d'ensemble de Covid-on-the-Web : pipeline, ressources, services et applications.

2.2 Construction d'un graphe de connaissances d'argumentation

Argumentative **C**linical **T**rial **A**nalysis (ACTA) [11] est un outil conçu pour analyser les composants argumentatifs des essais cliniques, s'appuyant sur la méthode PICO. [17] Développé à l'origine comme un outil de visualisation interactif pour aider les cliniciens dans l'analyse des essais cliniques, nous avons adapté ACTA pour annoter le corpus CORD-19. ACTA va bien au-delà de la simple recherche par mot-clés en extrayant la ou les affirmations (claims) principales énoncées dans un article, ainsi que les preuves (evidences) supportant cette affirmation, et les éléments PICO correspondants.

Dans le contexte des essais cliniques, une *affirmation* est une déclaration finale faite par l'auteur sur le résultat de l'étude. Elle décrit généralement la relation d'un nouveau traitement par rapport aux traitements existants. Par conséquent, une observation ou une mesure est une *preuve* qui soutient ou attaque un autre composant argumentatif. Les observations comprennent les effets secondaires et les résultats obtenus. Deux types de relations peuvent exister entre les composants argumentatifs. La relation dit d'*attaque* tient lorsqu'un composant contredit la proposition de la composante cible, ou déclare que les effets observés ne sont statistiquement pas significatifs. La relation dit de *support* s'applique à toutes les déclarations ou observations justifiant la proposition du composant cible.

Chaque résumé du corpus CORD-19 a été analysé par ACTA [18] et le résultat représenté en RDF afin de générer le graphe de connaissances d'argumentation CORD-19 (CORD19-AKG). Le pipeline se compose de quatre étapes : *(1)* la détection des composants argumentatifs, c.-à-d. détecter les affirmations et les preuves, *(2)* la prédiction des relations existant entre ces composants, *(3)* l'extraction des éléments PICO, et *(4)* la production de la représentation RDF des arguments et des éléments PICO.

Détection de composants argumentatifs. Il s'agit d'une tâche d'étiquetage séquentiel où, pour chaque mot, le modèle prédit si le mot fait partie d'un composant ou non. Nous associons l'architecture BERT [6] à un LSTM (un réseau récurrent à mémoire court et long terme) [17] et un champ aléatoire conditionnel (conditional random fields) pour effectuer de la classification des unités lexicales (tokens). Les poids de BERT sont initialisés à partir des poids de SciBERT [1], ce qui permet une meilleure représentation textuelle des articles scientifiques utilisés dans un corpus tel que CORD-19. Le modèle optimisé (fine-tuned) sur un jeu de données annoté avec des composants argumentatifs obtient une f_1-score de 0,90 [12].

Classification des relations. Les composants argumentatifs extraits à partir de l'étape précédente sont ensuite évalués conjointement pour définir leurs inter-relations. Déterminer quelles relations existent entre les composants est une tâche d'étiquetage séquentiel à trois classes, où la séquence est constituée d'une paire de composants, et où la tâche est d'apprendre la relation entre eux, c.-à-d. *support, attaque* ou *aucune relation*.

Le transformeur SciBERT est utilisé pour créer la représentation vectorielle du texte en entrée, auquel on ajoute une couche linéaire afin de déterminer les relations. Le modèle est optimisé sur un jeu de données comportant les relations d'argumentation dans le domaine médical et obtient une f_1-score de 0,68 [12].

Détection des éléments PICO. Nous utilisons la même architecture que pour la détection des composants. Le modèle est entraîné sur le corpus EBM-NLP [16] afin de prédire la/le population / patient / problème (P de PICO), l'intervention (I de PICO) [19] et les critères de jugement (O de PICO) pour une entrée donnée. L'évaluation de la détection des éléments PICO obtient une f_1-score de 0.734 [11]. Chaque composant argumentatif est annoté avec les éléments PICO qu'il contient. Pour faciliter les requêtes structurées, les éléments PICO sont rattachés à leurs concepts UMLS (Unified Medical Language System) à l'aide de ScispaCy [15].

Graphe de connaissances d'argumentation. Le *graphe de connaissances d'argumentation CORD-19* (CORD19-AKG) s'appuie sur l'Argument Model Ontology (AMO), [20]

17. PICO est un cadre utilisé pour répondre aux questions de soins de santé dans la pratique fondée sur des preuves. Il signifie : patients/population (P), intervention (I), control/comparison (C) et outcome (O).

18. https://github.com/Wimmics/CovidOnTheWeb/tree/master/src/acta

19. L'étiquette d'intervention (I de PICO) et de comparaison (C de PICO) sont traitées comme appartenant à une même classe.

20. http://purl.org/spar/amo/

Listing 2 – Exemple des composants argumentatifs et de leur relation.

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

<http://ns.inria.fr/covid19/arg/4f8...>
 a amo:Argument;
 schema:about <http://ns.inria.fr/covid19/4f8...>;
 amo:hasEvidence
 <http://ns.inria.fr/covid19/arg/4f8.../0>;
 amo:hasClaim <http://ns.inria.fr/covid19/arg/4f8.../6>.

<http://ns.inria.fr/covid19/arg/4f8.../0>
 a amo:Evidence, sioca:Justification, aif:I-node;
 prov:wasQuotedFrom <http://ns.inria.fr/covid19/4f8...>;
 aif:formDescription "17 patients discharged in
                      recovered condition...";
 sioca:supports
 <http://ns.inria.fr/covid19/arg/4f8.../6>;
 amo:proves <http://ns.inria.fr/covid19/arg/4f8.../6>.
```

le SIOC Argumentation Module (SIOCA)[21] et l'Argument Interchange Format.[22] Chaque argument identifié par ACTA est modélisé comme un `amo:Argument` dont les composants argumentatifs, affirmations et preuves, sont reliés. Les affirmations et les preuves sont elles-mêmes reliées par des relations de support ou de réfutation (avec respectivement les propriétés `sioca:supports`/`amo:proves` et `sioca:challenges`).

Le Listing 2 dresse un exemple des composants de ce graphe. Les éléments PICO (non illustrés dans le Listing 2) sont décrits comme des annotations des éléments argumentatifs dans lesquels ils ont été identifiés, de manière très similaire aux entités nommées (Listing 1). La différence est que le corps (body) de l'annotation contient les identifiants UMLS des concepts (CUI) et de types (TUI).

2.3 Génération du jeu de données

D'un point de vue technique, le corpus CORD-19 se compose d'un document JSON par article scientifique. La génération du jeu de données RDF Covid-on-the-Web implique donc deux étapes principales : (1) traiter chaque document du corpus pour en extraire les entités nommées et les arguments, et (2) traduire les résultats de ces deux traitements en un jeu de données RDF unifié et cohérent. L'ensemble du pipeline est décrit dans la Figure 1.

Extraction des entités nommées. Pour chaque article du corpus, DBpedia Spotlight, Entity-fishing et BioPortal Annotator produisent chacun un document JSON allant de 100 KB à 50 MB chacun. Ces documents ont été chargés dans une base de données MongoDB, et prétraités pour filtrer les données inutiles ou invalides (ex. les caractères non valides) ainsi que pour supprimer les entités nommées de moins de trois caractères. Ensuite, chaque document a été traduit en RDF tel que décrit dans la section 2.1 en utilisant Morph-xR2RML,[23] une implémentation pour MongoDB du langage de transformation xR2RML [13]. Les trois annotateurs sémantiques ont été déployés sur une Precision Tower 5810 équipée d'un CPU à 3,7 GHz et de 64 Go de RAM.

Pour que les fichiers générés par Annotator+ conservent une taille manipulable, nous avons désactivé les options relatives à la négation (`negation`), à la détection du patient impliqué dans une expression médicale (`experiencer`), à la temporalité (`temporality`), à la hiérarchie des concepts identifiés (`display_links`) et aux informations sur les vocabulaires requêtés (`display_context`). Nous avons activé l'option `longest_only`, ainsi que l'option de lemmatisation (`lemmatize`) pour améliorer les capacités de détection. MongoDB et Morph-xR2RML ont été déployés sur une autre machine équipée de 8 cœurs et de 48 Go de RAM.

Extraction du graphe d'arguments. Seuls les résumés de plus de dix mots ont été traités par ACTA pour garantir des résultats significatifs. Au total, 44 153 documents ont répondu à ce critère. ACTA a été déployé sur un nœud dual-Xeon de 2,8 GHz avec 96 Go de RAM.

Comme pour l'extraction des entités nommées, les documents JSON en sortie ont été chargés dans MongoDB et traduits dans la représentation RDF décrite dans la section 2.2 en utilisant Morph-xR2RML. La traduction en RDF a été effectuée sur la même machine que celle décrite ci-dessus (celle hébergeant MongoDB et Morph-xR2RML).

3 Publication et interrogation du jeu de données Covid-on-the-Web

Le jeu de données Covid-on-the-Web est composé de deux principaux graphes RDF, à savoir le graphe de connaissances des entités nommées CORD-19 et le graphe de connaissances d'argumentation CORD-19. Un troisième graphe décrit les métadonnées et le contenu des articles CORD-19. Le Tableau 1 synthétise la quantité de données en termes de documents JSON et de triples RDF produits.

Description du jeu de données. Conformément aux meilleures pratiques en matière de publication de données [7], nous fournissons une description détaillée du jeu de données Covid-on-the-Web lui-même. Celle-ci comprend notamment (1) des informations relatives aux licences, aux contributeurs et la provenance décrites avec DCAT,[24] et (2) aux vocabulaires, aux liens entre jeux de données et les informations pour accéder aux données avec VOID.[25]

Accessibilité des données. Le jeu de données RDF est identifié via un DOI, téléchargeable depuis la plateforme Zenodo et accessible au moyen d'un endpoint SPARQL public. Chaque URI peut être déférencée avec négociation de contenu.

Le dépôt Github du projet fournit une documentation exhaustive incluant des détails relatifs aux licences, aux représentations RDF, aux graphes nommés et aux ontologies chargées dans l'endpoint. Ces informations sont résumées dans le Tableau 2.

21. `http://rdfs.org/sioc/argument#`
22. `http://www.arg.dundee.ac.uk/aif#`
23. `https://github.com/frmichel/morph-xr2rml/`
24. `https://www.w3.org/TR/vocab-dcat/`
25. `https://www.w3.org/TR/void/`

TABLE 1 – Volume de données de Covid-on-the-Web.

Type de données	Données JSON	Ressources produites	Triplets RDF
Métadonnées sur les articles et le contenu	15 GB	n.a.	3.72 M
Graphe de Connaissances des Entités Nommées CORD-19 (CORD19-NEKG)			
ENs identifiées par DBpedia Spotlight (titres, résumés)	87 GB	4.1 M	65.4 M
ENs identifiées par Entity-fishing (titres, résumés, corps)	52 GB	66.1 M	1.16 G
ENs identifiées par BioPortal Annotator (titres, résumés)	378 GB	43 M	104.4 M
Graphe de Connaissances d'Argumentation CORD-19 (CORD19-AKG)			
Composants preuves / affirmations (résumés)	112 MB	119 k	7.47 M
Éléments PICO		515 k	
Données totales pour Covid-on-the-Web (en incluant les métadonnées sur les articles et le contenu)			
	532 GB	113 M entités nommées 119 k preuves/affirmations 515 k éléments PICO	1.36 G

TABLE 2 – Accessibilité de Covid-on-the-Web.

CovidOnTheWeb DOI	10.5281/zenodo.4247134
Données RDF	`https://doi.org/10.5281/zenodo.4247134`
Endpoint SPARQL	`https://covidontheweb.inria.fr/sparql`
Documentation	`https://github.com/Wimmics/CovidOnTheWeb`
Espace de nommage	`http://ns.inria.fr/covid19/`
CovidOnTheWeb URI	`http://ns.inria.fr/covid19/covidontheweb-1-2`

Reproductibilité. Conformément aux principes de la science ouverte, tous les scripts, fichiers de configuration et de traduction en RDF impliqués dans notre pipeline sont fournis dans le dépôt Github du projet selon les termes de la licence Apache 2.0, de sorte que n'importe qui peut relancer l'ensemble de la chaîne de traitement (de l'extraction des articles au chargement des fichiers RDF dans Virtuoso OS).

Licences. Les données produites pour Covid-on-the-Web sont dérivées du corpus CORD-19, et en tant que telles différentes licences s'appliquent à ces dernières. Les métadonnées sur les articles ainsi que le contenu des articles traduits en RDF de CORD-19 sont publiés sous les mêmes termes que la licence de CORD-19. [26]

Les résultats de l'extraction des articles, qu'il s'agisse des ENs (CORD19-NEKG) ou des composants argumentatifs (CORD19-AKG), sont publiés selon les termes de la licence d'attribution Open Data Commons 1.0 (ODC-By). [27]

Maintenance. Chaque semaine de nouvelles recherches sont publiées au sujet du Covid-19. La valeur de Covid-on-the-Web, ainsi que des autres jeux de données s'attaquant à cette problématique, réside dans leur habilité à pouvoir intégrer ces nouveaux résultats au fur et à mesure de leur publication. À cette fin, nous avons pris soin de produire un pipeline documenté et reproductible, et nous avons déjà effectué une telle mise à jour, validant ainsi notre démarche. À moyen terme, nous avons l'intention d'améliorer la fréquence des mises à jour en considérant à la fois *(1)* l'importance des mises à jour de CORD-19 (nombre de nouveaux articles), et *(2)* les besoins définis par l'expression de nouveaux scénarios d'application (voir Section 5). En outre, nous avons déployé un serveur permettant d'héberger un endpoint SPARQL qui bénéficie d'une infrastructure à haute disponibilité et d'un support 24 heures sur 24, 7 jours sur 7.

4 Visualisation et utilisations du jeu de données

Notre projet s'est également attaché à explorer les moyens de visualiser les données et d'interagir avec elles. Nous avons pour cela développé un outil appelé *Covid Linked Data Visualizer* [28] qui comprend une interface web hébergée sur un serveur node.js, un moteur de transformation basé sur Corese Semantic Web factory [4], et la bibliothèque graphique MGExplorer [3].

Par le biais de l'interface web, les utilisateurs peuvent utiliser des requêtes SPARQL prédéfinies ou écrire leurs propres requêtes. Des formulaires HTML servent à spécifier certains critères de recherche tels que la date de publication des articles. Par la suite, le moteur de transformation convertit les résultats des requêtes SPARQL dans le format JSON attendu par la bibliothèque graphique. L'exploration du graphe résultant est possible grâce à MGExplorer, une bibliothèque qui englobe un ensemble de techniques de visualisation spécialisées, chacune d'entre elles permettant de se concentrer sur un type de relation particulier.

La Figure 2 illustre certaines de ces techniques : le diagramme de graphe (gauche) montre une vue d'ensemble des nœuds et de leurs relations ; ClusterVis (en haut à droite) est une vue basée sur les clusters, qui permet de comparer les attributs des nœuds tout en préservant la représentation des relations entre eux ; IRIS (en bas à droite) est une vue égocentrique qui permet d'afficher tous les attributs et les relations d'un nœud donné. L'originalité de ces techniques de visualisation est d'offrir aux utilisateurs différents modes d'interaction qui peuvent les aider à explorer, classer et analyser l'importance des publications.

26. `https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/2020-03-13/COVID.DATA.LIC.AGMT.pdf`

27. `http://opendatacommons.org/licenses/by/1.0/`

28. `http://covid19.i3s.unice.fr:8080`

Lors d'une réunion avec des organismes liés au domaine de la santé et de la recherche médicale (Inserm et INCa), un expert nous a indiqué un exemple de requête que les chercheurs aimeraient faire sur un tel jeu de données : "Identifier les articles qui mentionnent à la fois un type de cancer et un virus de la famille des coronavirus". En prenant en considération cette requête, nous avons utilisé Covid Linked Data Visualizer et affiché les résultats à l'aide de la bibliothèque MGExplorer (Figure 2).

Nous avons également créé plusieurs Notebooks Jupyter, Python et R [29] pour montrer que ces résultats peuvent être convertis en Dataframes (des structures de données tabulaires utilisées en analyse des données) afin de procéder à de la fouille de données (Figure 3).

5 Impact potentiel et exploitation

À notre connaissance, le jeu de données Covid-on-the-Web est le premier à intégrer dans un seul et même ensemble cohérent des ENs, arguments et éléments PICO . Nous pensons qu'il pourra servir de base pour des applications du Web sémantique, pour des algorithmes d'analyse comparative ou pour des défis.

Les ressources et les services que nous proposons, liés à la littérature concernant la COVID-19, sont intéressants pour les organismes et les instituts de santé puisqu'ils permettent d'extraire et d'analyser efficacement les informations sur une maladie encore relativement inconnue et pour laquelle la recherche est en constante évolution. Dans une certaine mesure, il est possible de croiser les connaissances pour mieux appréhender ce sujet et, en particulier, pour initier des recherches sur des voies inexplorées. Nous espérons également que l'ouverture des données et du code permettra aux contributeurs de faire progresser l'état actuel des connaissances sur cette maladie dont l'impact sanitaire est mondial.

En plus d'être interopérables avec les graphes de connaissances majeurs utilisés au sein de la communauté du Web Sémantique, les visualisations que nous offrons au moyen de MGExplorer et de Notebooks Jupyter montrent le potentiel de ces technologies dans d'autres domaines, à titre d'exemple, dans les domaines biomédicaux et médicaux.

Documentation / Tutoriels. Nous avons conservé les documents méthodologiques que nous avons suivis afin de pouvoir justifier nos choix de conception : La documentation technique des algorithmes et des représentations RDF, [30] les meilleures pratiques dans l'élaboration et publication des données (FAIR, Cool URIs, données liées à cinq étoiles, etc.) et des documentations destinées aux utilisateurs finaux (par exemple, les Notebooks Jupyter).

Scénarios, modèles d'utilisateurs et requêtes types. Nos ressources sont basées sur des outils génériques que nous avons adaptés à la problématique de la COVID-19. En adoptant une approche orientée utilisateur, nous les avons conçues selon trois principaux scénarios identifiés par une analyse des besoins des instituts biomédicaux avec lesquelles nous collaborons : *(Scénario 1)* Aider les cliniciens à obtenir des graphes argumentatifs pour analyser les essais cliniques et prendre des décisions fondées sur des données ; *(Scénario 2)* Aider les médecins en milieu hospitalier à collecter les valeurs biologiques (par exemple, le cholestérol) à partir d'articles scientifiques, afin de déterminer si leurs patients sont dans les normes ou non ; *(Scénario 3)* Aider les chefs de mission d'un institut du cancer à identifier les articles scientifiques traitant du cancer et des coronavirus afin d'élaborer des programmes de recherche pour étudier les liens entre eux.

La généricité des outils que nous avons développés nous permet de les appliquer à un panel plus large de scénarios, et nos partenaires dans le domaine biomédical nous incitent déjà à réfléchir à des scénarios liés à d'autres questions que la COVID-19.

Outre les scénarios décrits ci-dessus, nous établissons également des modèles d'utilisateurs représentatifs (sous la forme de personas), dont le but est de nous aider à identifier les besoins, l'expérience, les comportements et les objectifs de nos utilisateurs.

Nous avons également reçu diverses demandes des utilisateurs potentiels que nous avons interrogés. Ces demandes servent à préciser et à tester notre graphe de connaissances et nos services. À des fins de généricité, nous avons élaboré une typologie à partir des demandes collectées, en utilisant des dimensions telles que : demande prospective vs. rétrospective ou demande descriptive (demande de description) vs. explicatives (demande d'explication) vs. argumentatives (demande d'argumentation). Voici des exemples de ces requêtes :

(Demandes descriptives prospectives) Quels types de cancers sont susceptibles d'apparaître chez les victimes de la COVID-19 au cours des prochaines années ? Chez quelles catégories de patients ? Etc.

(Demandes rétrospectives descriptives) Quels types de cancers sont apparus chez les victimes de [SARSCoV1 | MERS-CoV] au cours des [2|3|n] années suivantes ? Quel était le taux d'occurrence ? Chez quels types de patients ? Etc. Quelles sont les différentes séquelles liées aux coronavirus ? Quels sont les patients guéris de la COVID-19 qui ont une fibrose pulmonaire ?

(Demandes rétrospectives explicatives) Le [SARS-CoV1 | MERS-CoV] peut-il induire le cancer ? La [malignité | progression du cancer] est-elle directement induite par une infection au coronavirus ? Ou était-elle indirectement causée par les [inflammations | modifications métaboliques] liées à une infection ? Quelles séquelles liées aux coronavirus sont responsables du plus fort potentiel de malignité ?

(Demande rétrospective argumentatives) Quelles sont les preuves que le [SARS-CoV1 | MERS-CoV] provoque le cancer ? Quelles expériences ont démontré que la fibrose pulmonaire observée chez les patients guéris de la COVID-19 était causée par la COVID-19 ?

Ces requêtes sont une brève illustration d'une liste réelle (mais non exhaustive) de questions avancées par les utilisateurs. Certaines questions peuvent trouver une réponse en

29. https://github.com/Wimmics/covidontheweb/tree/master/notebooks

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

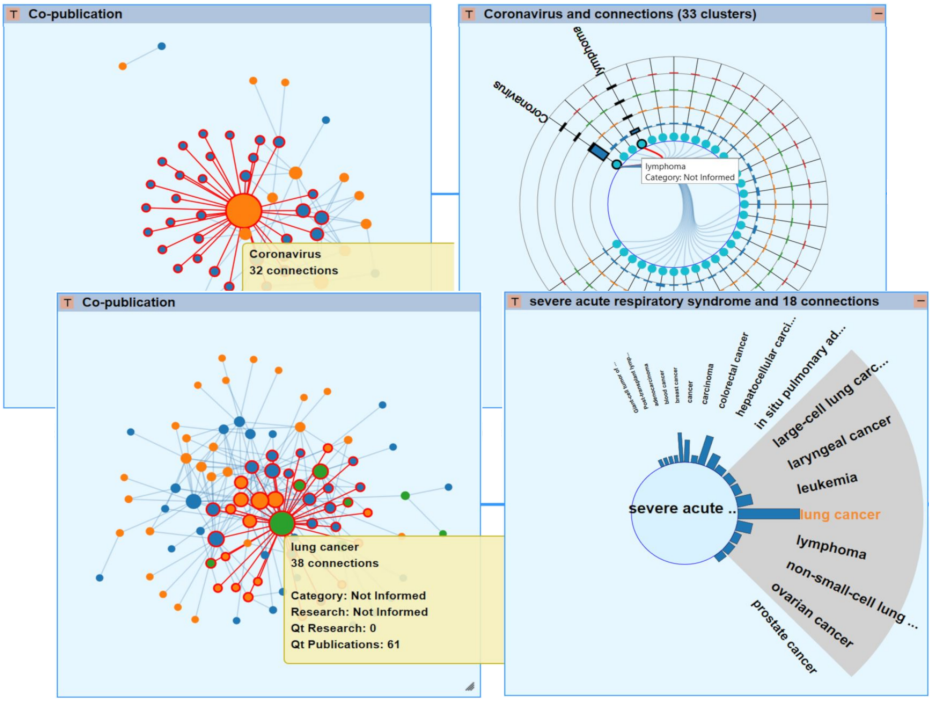

FIGURE 2 – Covid Linked Data Visualizer : visualisation des articles qui mentionnent à la fois un type de cancer (points bleus) et un virus de la famille des coronavirus (points orange).

montrant la corrélation entre les composants (par exemple, les types de cancer), d'autres nécessitent de représenter des tendances (par exemple, le cancer susceptible de se produire au cours des prochaines années) et l'analyse d'attributs spécifiques (par exemple, des détails sur les changements métaboliques causés par la COVID-19). La réponse à ces questions complexes requiert l'exploration du corpus CORD-19, et pour cela nous offrons une variété d'outils d'analyse et de visualisation. Ces requêtes et la typologie générique seront réutilisées dans d'autres extensions ainsi que d'autres projets.

Le Covid Linked Data Visualizer (présenté dans la Section 4) permet l'exploration visuelle du jeu de données Covid-on-the-Web. Les utilisateurs peuvent inspecter les éléments du graphe généré par une requête SPARQL (en positionnant la souris sur les éléments) ou explorer le graphe de façon itérative en chaînant les visualisations et en utilisant l'une des techniques d'interaction disponibles (que ce soit par IRIS, ClusterVis, etc.). Ces techniques de visualisation sont destinées à aider les utilisateurs à comprendre les relations présentes au sein des résultats. Par exemple, les utilisateurs peuvent lancer une requête pour visualiser un réseau de co-auteurs ; puis se servir de IRIS et ClusterVis pour comprendre qui collabore ensemble et sur quelles thématiques. Ils peuvent également lancer une recherche pour trouver des articles mentionnant la COVID-19 et divers types de cancer. Enfin, le mode avancé permet d'ajouter de nouvelles requêtes SPARQL mettant en œuvre d'autres chaînes d'exploration de données.

6 Conclusion et perspectives

Nous avons décrit dans cet article les données et logiciels déployés par le projet Covid-on-the-Web. Nous avons adapté et combiné des outils pour traiter, analyser et enrichir le corpus CORD-19 afin de permettre aux chercheurs dans le domaine biomédical d'accéder plus aisément à la littérature relative à la COVID-19, de l'interroger et de lui donner sens.

Nous avons conçu et publié un graphe de connaissances des données liées décrivant les entités nommées mentionnées dans les articles de CORD-19 et les graphes d'argumentation qu'ils incluent. Nous avons également publié le pipeline mis en place pour générer ce graphe de connaissances, afin de *(1)* continuer à l'enrichir et de *(2)* faciliter la réutilisation et l'adaptation du jeu de données et du pipeline.

Au-delà de ce graphe de connaissances, nous avons également développé, adapté et déployé plusieurs outils fournissant des visualisations de données liées, des méthodes d'exploration et des Notebooks pour les spécialistes dans les sciences des données. Par nos interactions avec des instituts dans le domaine de la santé et de la recherche médicale (entretiens, observations, tests d'utilisateurs), nous continuons de nous assurer que notre approche est guidée par et alignée sur les besoins de la communauté biomédicale. Nous avons montré que notre jeu de données permet d'effectuer des recherches documentaires et fournir des visualisations adaptées aux besoins des experts. De plus, notre démarche, dès ses prémices, s'est attachée à répondre aux objectifs de la science ouverte et reproductible ainsi qu'aux principes FAIR.

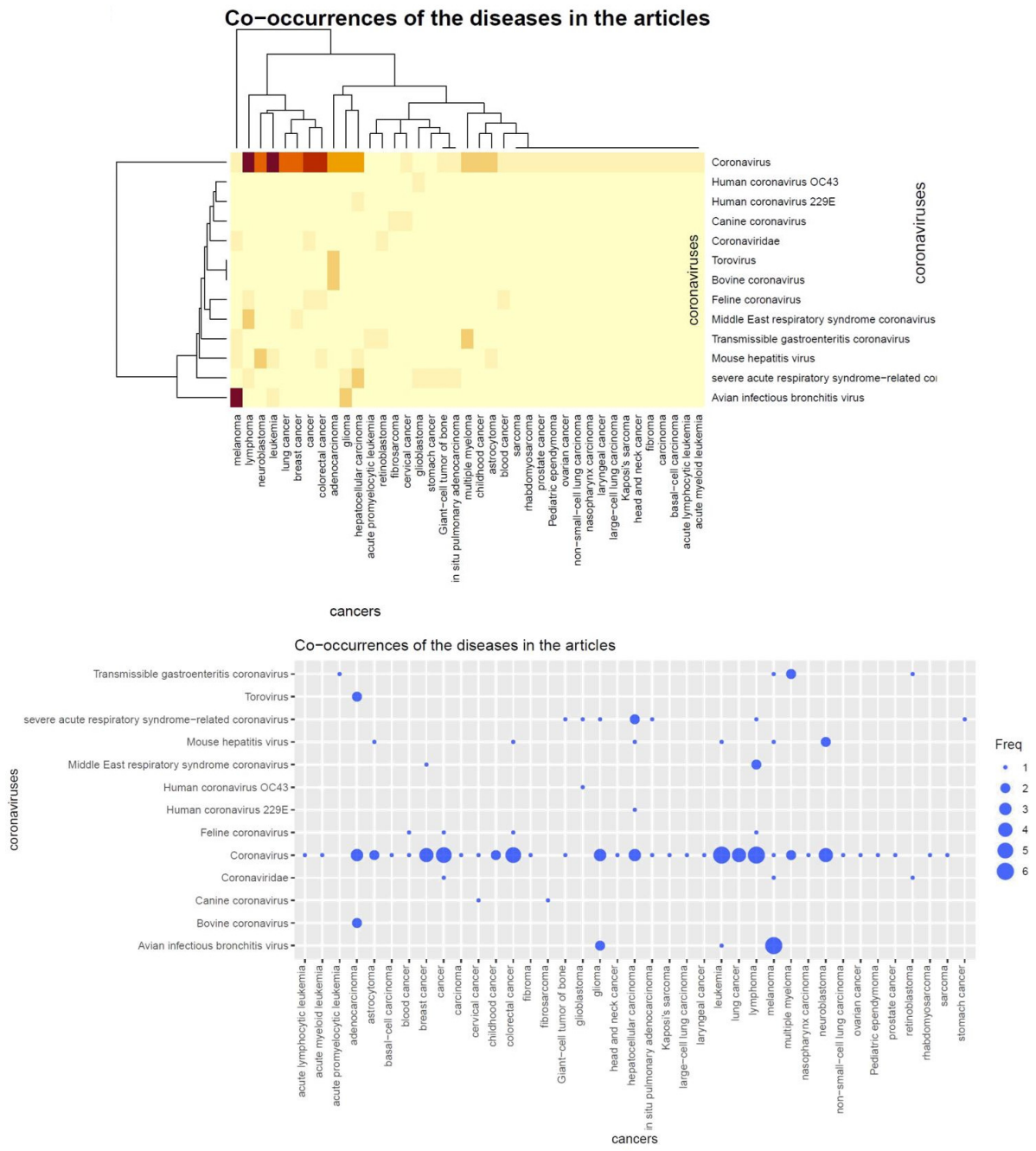

FIGURE 3 – Deux représentations différentes sous Jupyter Notebook du nombre d'articles qui co-mentionnent les types de cancer et les virus de la famille des coronavirus.

Depuis l'émergence de la COVID-19, le rythme effréné auquel les nouvelles recherches ont été publiées et les bases de connaissances ont évolué pose des problèmes critiques. Par exemple, de nouvelles versions de CORD-19 étaient publiées chaque semaine (désormais ce rythme peut être journalier), ce qui remet en question la capacité à suivre les dernières avancées. Par ailleurs, l'extraction et la désambiguïsation des ENs sur notre première version du jeu de données avaient été réalisées à l'aide de modèles pré-entraînés produits avant la pandémie, donc avant même la création de l'entité SARS-Cov-2 dans Wikidata. Par conséquent, à moyen terme, nous avons l'intention de nous engager dans un objectif de maintenance pérenne visant à ingérer régulièrement de nouvelles données, suivre l'évolution des connaissances et mettre régulièrement à jour nos extracteurs. Étant donné qu'il n'existe pas de jeu de données de référence de CORD-19 qui aurait été manuellement annoté et qui pourrait donc servir de référence (gold standard), il est difficile d'évaluer la qualité des modèles utilisés pour extraire les ENs et les structures argumentatives. Pour pallier ce problème, nous travaillons sur la mise en œuvre de curation de contenu (data curation), et la découverte auto-

matisée de motifs et de règles d'association qui pourraient être utilisés pour détecter les erreurs dans l'extraction des ENs, permettant ainsi de proposer un contrôle qualité des données.

Références

[1] I. Beltagy, K. Lo, and A. Cohan. SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, 2019.

[2] M. Bersanelli. Controversies about COVID-19 and anticancer treatment with immune checkpoint inhibitors. *Immunotherapy*, 12(5) :269–273, April 2020.

[3] R. Cava, C. Freitas, and M. Winckler. Clustervis : visualizing nodes attributes in multivariate graphs. In *Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, April 3-7, 2017*, pages 174–179. ACM, 2017.

[4] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the semantic web with Corese search engine. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, volume 16, page 705, Valencia, Spain, 2004.

[5] J. Daiber, M. Jakob, C. Hokamp, and P. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 121–124, 2013.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 4171–4186, 2019.

[7] B. Farias Lóscio, C. Burle, and N. Calegari. Data on the Web Best Practices. *W3C Recommandation*, 2017.

[8] R. Gazzotti, C. Faron-Zucker, F. Gandon, V. Lacroix-Hugues, and D. Darmon. Injecting domain knowledge in electronic medical records to improve hospitalization prediction. In *The Semantic Web - 16th European Conference, ESWC, Portorož, Slovenia, June 2-6, 2019, Proceedings*, volume 11503 of *Lecture Notes in Computer Science*, pages 116–130. Springer, 2019.

[9] R. Gazzotti, C. Faron-Zucker, F. Gandon, V. Lacroix-Hugues, and D. Darmon. Injection of automatically selected DBpedia subjects in electronic medical records to boost hospitalization prediction. In *SAC '20 : The 35th ACM/SIGAPP Symposium on Applied Computing, online event, March 30 - April 3, 2020*, pages 2013–2020. ACM, 2020.

[10] C. Jonquet, N. Shah, and M. Musen. The open biomedical annotator. *Summit on translational bioinformatics*, 2009 :56, 2009.

[11] T. Mayer, E. Cabrio, and S. Villata. ACTA a tool for argumentative clinical trial analysis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6551–6553, 2019.

[12] T. Mayer, E. Cabrio, and S. Villata. Transformer-based argument mining for healthcare applications. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, 2020.

[13] F. Michel, L. Djimenou, C. Faron-Zucker, and J. Montagnat. Translation of Relational and Non-Relational Databases into RDF with xR2RML. In *Proceeding of the 11th International Conference on Web Information Systems and Technologies (WebIST)*, pages 443–454, Lisbon, Portugal, 2015.

[14] F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby, R. Gazzotti, A. Giboin, S. Marro, T. Mayer, M. Simon, S. Villata, and M. Winckler. Covid-on-the-web : Knowledge graph and services to advance COVID-19 research. In *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, volume 12507 of *Lecture Notes in Computer Science*, pages 294–310. Springer, 2020.

[15] M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy : Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[16] B. Nye, J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, and B. Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proc. 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 197–207, 2018.

[17] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv :1402.1128*, 2014.

[18] L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. Weld, O. Etzioni, and S. Kohlmeier. Cord-19 : The covid-19 open research dataset. *ArXiv*, abs/2004.10706, 2020.