Revista Española de Documentación Científica
43(4), octubre-diciembre 2020, e275
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2020.4.1794

ESTUDIOS / *RESEARCH STUDIES*

Sistemas de recuperación de información implementados a partir de CORD-19: herramientas clave en la gestión de la información sobre COVID-19

Rosana López Carreño*, Francisco Javier Martínez Méndez*

Departamento de Información y Documentación, Universidad de Murcia
Correo-e: rosanalc@um.es | ORCID iD: https://orcid.org/0000-0002-2097-9389
Correo-e: javima@um.es | ORCID iD: https://orcid.org/0000-0003-1098-9361

Recibido: 20-06-20; 2ª versión: 17-09-20; Aceptado: 17-09-20.

Cómo citar este artículo/Citation: López Carreño, R.; Martínez Méndez, F. J. (2020). Sistemas de recuperación de información implementados a partir de CORD-19: herramientas clave en la gestión de la información sobre COVID-19. *Revista Española de Documentación Científica*, 43 (4), e275. https://doi.org/10.3989/redc.2020.4.1794

Resumen: La investigación sobre el coronavirus ha generado una producción de documentos científicos extraordinaria. Su tratamiento y asimilación por parte de la comunidad científica ha necesitado de la ayuda de sistemas de recuperación de información diseñados específicamente. Algunas de las principales instituciones mundiales dedicadas a la lucha contra la pandemia han desarrollado el conjunto de datos CORD-19 que destaca sobre otros proyectos de similar naturaleza. Los documentos recopilados en esta fuente han sido procesados por distintas herramientas de recuperación de información, a veces prototipos o sistemas que ya estaban implementados. Se ha analizado la tipología y características principales de estos sistemas concluyendo que hay tres grandes categorías no excluyentes entre ellas: búsqueda terminológica, visualización de información y procesamiento de lenguaje natural. Destaca enormemente que la gran mayoría de ellos emplean preferentemente tecnologías de búsqueda semántica con el objeto de facilitar la adquisición de conocimiento s los investigadores y ayudarlas en su ingente tarea. La crisis provocada por la pandemia ha sido aprovechada por los buscadores semánticos para encontrar su sitio.

Palabras clave: conjuntos de datos; COVID-19; Sistemas de Recuperación de Información; gestión de información; CORD-19.

Information retrieval systems implemented from CORD-19: key tools in managing information about the COVID-19

Abstract: Research on the coronavirus has generated an extraordinary production of scientific documents. Their treatment and assimilation by the scientific community has required the help of specifically designed information retrieval systems. Some of the world's leading institutions involved in the fight against the pandemic have developed the CORD-19 dataset that stands out from other projects of a similar nature. The documents collected in this source have been processed by various information retrieval tools, sometimes prototypes or previously implemented systems. The typology and main characteristics of these systems have been analysed, concluding that there are three main non-exclusive categories among them: terminological search, information visualisation and natural language processing. It should be noted that most of them use semantic search technologies in order to facilitate the acquisition of knowledge by researchers and to help them in their enormous task. The crisis caused by the pandemic has been taken advantage of by semantic search engines to find their site.

Keywords: datasets; COVID-19; Information Retrieval Systems; information management; CORD-19.

1. INTRODUCCIÓN

El mundo está viviendo los efectos de una pandemia de origen indefinido y alcance aún por determinar. Sus consecuencias y los esfuerzos que se están llevando a cabo para frenar sus efectos son profundamente sociales. Estos esfuerzos involucran la coordinación de grandes comunidades de científicos, gestores políticos y administrativos y de la ciudadanía (Adams y Light, 2020). Es importante comprender las dimensiones sociales que vertebran este empuje que está modificando el desarrollo de la investigación científica en un espacio muy breve de tiempo.

La descripción por medio de metadatos de la información científica y técnica, la normalización terminológica y conceptual a través de tesauros y lenguajes facetados propios de las Ciencias de la Salud, la minería de datos y el desarrollo de sistemas de información interoperables (a través de servicios REST y APIs, entre otros), ha provocado la generación *ad hoc* de una amplia variedad de fuentes de información específicas en torno al Coronavirus, no solo de artículos científicos sino también de casos clínicos, datos epidemiológicos, evidencias o patentes. Esta reacción en cadena de revistas biomédicas, editoriales, universidades, instituciones de investigación y empresas de desarrollo de inteligencia artificial (IA) ha derivado en una propagación de la información científica sobre el COVID-19 en paralelo a la del propio virus situando a los profesionales de la información en el centro de la pandemia informativa (Torres-Salinas 2020).

Con la intención de frenar esta crisis sanitaria global, revistas biomédicas del prestigio de *New England Journal of Medicine*, *JAMA*, *Lancet*, *Nature*, *Science*, *Cell* o *British Medical Journal*, entre otras, disponen en sus sedes webs de material bibliográfico propio publicado en acceso abierto. Las principales editoriales del ámbito de la salud también han creado espacios de información selectiva con ecuaciones de búsqueda predefinidas sobre los principales tópicos de la investigación relacionada con la pandemia y permiten aplicar varios conjuntos de filtros a los resultados. Ejemplos de estos espacios son Cambridge Coronavirus Free Access Collection, EBSCO Covid-19, Elsevier Coronavirus Research Repository, Emerald COVID 19, SAGE, Wiley COVID-19, Oxford University Press o Cochrane Coronavirus (COVID-19). Del mismo modo, algunos repositorios temáticos del campo de las Ciencias de la Salud también han seleccionado material bibliográfico (artículos y preprints) que tratan sobre el tema en cuestión, como por ejemplo ArXiv, MedRxiv, Biorxiv, Pubmed o la colección especial COVID-19 del CSIC (Vicepresidencia de Investigación Científica y Técnica).

Las grandes plataformas bibliográficas y buscadores académico-científicos también han dispuesto material bibliográfico sobre la COVID-19 aplicando ecuaciones de búsqueda predeterminadas y filtros, así como conjuntos de datos estructurados: Dimensions (https://covid-19.dimensions.ai), Kaggle (https://www.kaggle.com/covid19), Google Académico, Microsoft Academic y Semantic Scholar. También hay sistemas de búsqueda de información sobre patentes como Lens (https://about.lens.org/covid-19/) o casos clínicos en Kahun (https://coronavirus.kahun.com), además de lo dispuesto por los buscadores específicos de conjuntos de datos, como son, por ejemplo, Data World, Microsoft Research Open Data y Google Dataset Search. Asimismo, como era de esperar, las autoridades sanitarias en unión con universidades, centros de investigación y las sociedades más relevantes en este campo científico han desarrollado servicios informativos sobre el Coronavirus: National Institutes of Health (NIH), Centers for Disease Control and Prevention (CDC), Organización Mundial de la Salud (OMS/WHO), Utrecht University, Johns Hopkins University, American Society for Microbiology y Global Rheumatology Alliance.

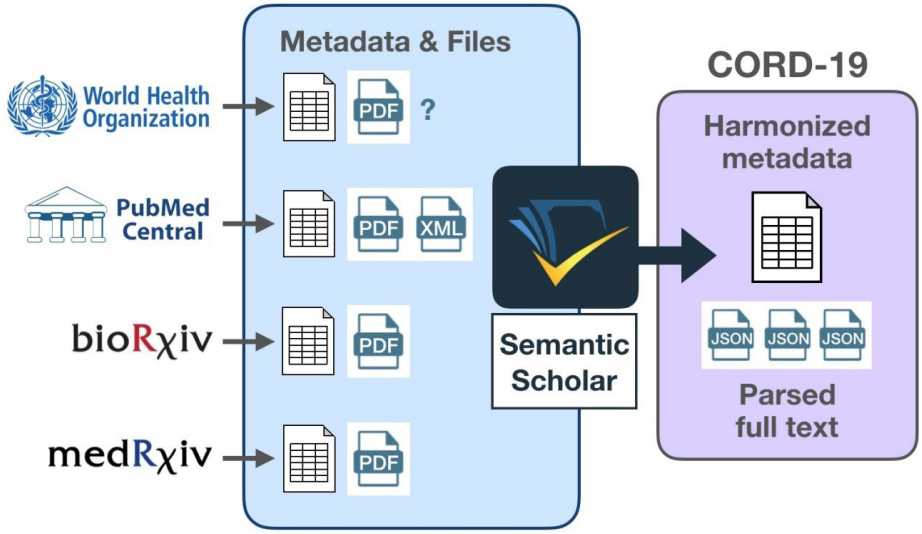
La emergencia informativa sobre el virus no solo ha recaído en el ámbito científico sino también en el ámbito político y social, lo que justifica la proliferación de conjuntos de datos de distinto tipo, estructura, formato y cobertura produciendo una saturación en su identificación y gestión. Desde nuestro punto de vista, esto justifica la necesidad de abordar una diferenciación entre los mismos, clasificando los conjuntos de datos en tres tipos: estadísticos o epidemiológicos, servicios de búsqueda terminológicos o semánticos y conjuntos de datos bibliográficos. Los conjuntos de datos estadísticos o epidemiológicos se nutren de las series de datos aportadas por los estados y por las instituciones internacionales del ámbito de la salud. Esta información, a veces, no se dispone en formatos y estructuras limpias para su reutilización y esto ha hecho preciso del desarrollo de herramientas de visualización y actualización, que ayuden no solo a la toma de decisiones sino a la información social, creándose infografías y visualizaciones de datos útiles y provechosas no solo para médicos e investigadores sino también para los gestores de información y los profesionales de la comunicación (Callaghan, 2020). En este conjunto de recursos destaca 'Information is Beautiful', (https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack) o el Centro de Recursos de Coronavirus de la Universidad John Hopkins (https://coronavirus.jhu.edu/map.html) cuyo mapa está enlazado a más de 200.000 sitios web y es el referente informativo de los medios de comunicación.

En el ámbito de las ciencias de la salud, la normalización terminológica y el uso de lenguajes controlados como tesauros y listas de términos encabezamientos de materia, como MESH (Medical Subject Headings) o DeCS (Descriptores en Ciencias de la Salud) en las revistas biomédicas, tienen un valor equiparable a la propia producción científica debido a que aumenta la eficiencia y precisión en la recuperación de información (Baumann, 2016). Por ello, los conjuntos de datos terminológicos (que pueden denominarse también como conceptuales o facetados) representan un eslabón fundamental en la gestión de información, como es el caso de CORD-19, y sus corpus documentales son la base para el desarrollo de motores de búsqueda basados en conceptos y mapas de relaciones entre ellos para ayudar a los usuarios en la búsqueda pertinente de información sobre el Coronavirus.

La creación de los conjuntos de datos abiertos generados en torno al Coronavirus puestos a disposición de la comunidad científica ha sido posible gracias a la implementación de muchas aplicaciones ya existentes para otros propósitos, permitiendo la explotación de la información sobre la pandemia y la posterior representación de su contenido (Fernández-Sellers y otros, 2019). Los conjuntos de datos bibliográficos recogen datos estructurados de investigación relativos a publicaciones científicas y de patentes cuyo contenido es agregado desde otros conjuntos de datos, repositorios o gestores de contenido bibliográfico. La diferencia entre unos y otros la marca el volumen de datos que contenga, destacando CORD-19 (COVID-19 Open Research Dataset, https://cord-19.apps.allenai.org) desarrollado desde el 16 de marzo de 2020 por The Allen Institute for Artificial Intelligence (AI2, centro de investigación impulsado por Paul Allen, cofundador de Microsoft) en colaboración con la Oficina de Política de Ciencia y Tecnología de Estados Unidos, la Biblioteca Nacional de Medicina (NLM), la iniciativa Chan Zuckerberg (promovida por Mark Zuckerberg, fundador de la red social Facebook y su esposa), Microsoft Research y el conjunto de datos Kaggle, bajo la coordinación del Centro de Seguridad y Tecnologías Emergentes de la Universidad de Georgetown (Wang y otros, 2020a).

Esta fuente agrega información semanalmente desde los repositorios PubMed, BioRxiv, MedRxiv y WHO/OMS. Existe una gran sinergia entre CORD-19 y el buscador semántico académico Semantic Scholar (proyecto también desarrollado en el instituto AI2, lanzado en el año 2015 aunque su verdadero punto de inflexión fue su asociación en 2018 con el buscador Microsoft Academic) que incluso permite la descarga del conjunto de datos en su página principal y que le ha ayudado a convertirse en el referente informativo para los investigadores durante esta pandemia (por el contrario, Google Scholar, motor mucho más desarrollado, no ha llevado a cabo algo parecido). Desde su lanzamiento, CORD-19 se ha descargado más de 75,000 veces y ha servido como base de muchos sistemas de minería y de descubrimiento de texto (Wang, y otros 2020a). En su apartado bibliográfico tiene 65.000 artículos científicos disponibles, muchos de ellos en versión preprint y más de la mitad en acceso abierto.

Figura 1. Fuentes de información de CORD-19



Fuente: Wang, et al. 2020.

Todo ello convierte a CORD-19 en uno de los conjuntos de datos más utilizados y referenciados entre las fuentes de información científicas creadas ad hoc sobre el Coronavirus. El objeto de su puesta en marcha es movilizar a los investigadores a aplicar los avances recientes en el procesamiento del lenguaje natural para generar nuevas ideas en apoyo de la lucha contra esta enfermedad infecciosa (Colavizza et at. 2020). Estos mismos autores describen a nivel cienciométrico el contenido de este conjunto de datos, obteniendo como conclusiones:

1. Las publicaciones CORD-19 están relacionadas más ampliamente con la investigación médica sobre virus, de los cuales COVID-19 y coronavirus son una parte.
2. Los temas dominantes en CORD-19 incluyen investigación sobre salud pública y epidemias; biología molecular; coronavirus, influenza y otras familias de virus; inmunología y antivirales; metodología (prueba, diagnóstico, ensayos).
3. La intensidad del tema en el tiempo está lejos de ser uniforme, lo que demuestra en particular que la investigación de coronavirus ha seguido brotes conocidos (SARS, MERS, COVID-19) y que hasta 2020 esta investigación representaba solo una pequeña porción de CORD-19.
4. Los grupos de redes de citas muestran una relativa cohesión de CORD-19, confirmando la amplia cobertura del conjunto de datos Parece haber dos grupos de citas prominentes: la investigación sobre coronavirus específicos, con un enfoque de salud pública y epidemiológico y otro con un enfoque de biología molecular.
5. Las almétricas certifican que el brote actual de SARS-CoV-2 domina la atención de las redes sociales, en particular de Twitter, destacando el interés público por los resultados científicos durante esta pandemia.

El conjunto de datos de la Universidad Johns Hopkins es la fuente de información estadística (epidemiológica) de referencia para consultar datos sobre el desarrollo de la pandemia. De forma paralela, y al mismo tiempo, CORD-19 se ha convertido en la principal fuente de referencia para la investigación, especialmente para acceder a los resultados de la investigación: fundamentalmente artículos. El esfuerzo desarrollado por la comunidad científica no tiene precedentes en volumen de producción y en la velocidad de transmisión: por ejemplo, Pubmed ha añadido publicaciones relacionadas diariamente desde primero de enero con un pico de 300 artículos en un solo día (Kousha y Thelwall, 2020). El volumen de información a manejar es ingente, el Big Data ayuda a los virólogos y a otros expertos en el manejo de la información estadística y en la identificación de posibles patrones de comportamiento en la evolución de la pandemia. En el otro lado, la sobrecarga informativa producida por la vasta producción científica será, en breve espacio de tiempo si no lo es ya, más un problema que una ayuda si no se disponen de herramientas informáticas que ayuden en la gestión y posterior recuperación de la información.

La investigación en la búsqueda de tratamientos eficaces contra el Coronavirus y de la anhelada vacuna ha movilizado investigadores en múltiples campos científicos, y los especialistas en recuperación de información, lejos de quedarse al margen, han ayudado de forma considerable en esta tarea: los cambios en la conducta y difusión de la ciencia crean desafíos para la recuperación de la información, el campo científico detrás de los motores de búsqueda (Roberts et al., 2020). La comunidad TREC-COVID surge con el objetivo de reunir equipos de investigación en recuperación de información para evaluar motores de búsqueda en tareas específicas e intentar reflejar los principales temas de interés de los usuarios de internet durante la pandemia y, al mismo tiempo, señala los principales documentos de trabajo sobre la COVID-19, sus síntomas, propagación, factores de riesgo y tratamientos (Dousset y Mothe, 2020).

En el campo de los buscadores web, una división muy simple establecería dos categorías: los terminológicos (en este grupo estarían los buscadores convencionales basados en el modelo vectorial y que alinean la respuesta en función de la similitud entre los términos de búsqueda y las palabras incluidas en los documentos) y los semánticos (que intentan llevar a cabo las búsquedas algo más en contexto y explorar las asociaciones de conceptos en la recuperación de la información). En el momento actual, en las circunstancias excepcionales provocadas por la pandemia, resulta complicado que los sistemas de búsqueda convencionales aporten soluciones rápidas y consistentes porque las funciones de similitud no permiten distinguir entre unos documentos de temática tan similar y no se dispone de la ayuda del factor de impacto y de las citas. Esto ha hecho rebrotar la importancia de la descripción de los documentos y conjuntos de datos en los repositorios científicos y también ha derivado en una apuesta clara y decidida de empresas e instituciones que trabajan con inteligencia artificial y minería de datos en la creación de herramientas de búsqueda específicas que apoyen a la investigación sobre el Coronavirus. Esta tremenda crisis

sanitaria les ha permitido mostrar prototipos y desarrollos más avanzados que quizá no habían tenido la suficiente audiencia e interés hasta ahora. Ante la extensa producción científica (gran parte de ella no ha sido revisada por pares porque no ha habido tiempo), los investigadores precisan de un nuevo paradigma para la recuperación de información ya que están obligados a filtrar entre una inmensa plétora de resultados y no disponen de las herramientas precisas para ello. De hecho, los motores de búsqueda tradicionalmente empleados en ciencias de la salud (PubMed, por ejemplo) están diseñados para la recuperación de documentos y no permiten la recuperación por medio de expresiones de búsqueda literales (o por frase exacta) posibilidad muy interesante en este entorno porque pueden servir como evidencias textuales que resultan ser clave para tareas como la generación de hipótesis y la validación de nuevos hallazgos (Wang y otros, 2020b).

El reto asumido por los desarrolladores de estos sistemas de búsqueda es tremendo, no solo por el inmenso número de documentos a procesar ni por el hecho de trabajar, al mismo tiempo, con contribuciones científicas revisadas, preprints y una heterogénea documentación de fuentes oficiales de naturaleza muy cambiante. El problema principal a resolver es más de fundamentos: estos cambios en la conducta y difusión de la ciencia crean desafíos para la recuperación de información. La recuperación de información persigue buscar rápidamente a través de una gran colección de documentos (corpus) para encontrar información relevante que satisfaga una necesidad informativa. Los objetivos biomédicos y de investigación en salud van, desde la promoción del descubrimiento científico hasta el apoyo a la toma de decisiones clínicas para abordar las necesidades de salud de los ciudadanos y combatir la información errónea. Todos estos son, por supuesto, altamente relevantes en una pandemia (Roberts y otros, 2020). Armonizar las prestaciones de los sistemas de recuperación de información con las necesidades de los investigadores y profesionales de ciencias de la salud es un verdadero desafío. Y comprobar cómo se ha llevado a cabo esta tarea constituye el principal objeto de este trabajo, junto con la identificación y clasificación de los desarrollos más relevantes en el ámbito de la recuperación de información puestos en marcha para ayudar en la búsqueda de la ansiada vacuna y/o tratamientos contra la enfermedad.

2. METODOLOGÍA

Para abordar este estudio se localizaron a través de los buscadores de conjuntos de datos Data World, Kaggle, Microsoft Research Open Data y Google Dataset Search, aquellos que agregan contenido desde CORD-19. También se seleccionaron motores de búsqueda o herramientas de recuperación de información que usan ese corpus documental. Además se consultó CORD-19 desde el sistema Semantic Scholar (https://www.semanticscholar.org/cord19) que indica las fuentes y recursos desarrollados bajo este conjunto de datos, así como su foro donde se van indicando los últimos desarrollos (https://discourse.cord-19.semanticscholar.org/t/cord-19-demos-and-resources/132).

Una vez identificadas las fuentes creadas de forma específica para gestionar información sobre la pandemia basadas en CORD-19, se procedió a categorizarlas como conjunto de datos o motor de búsqueda. Este segundo subconjunto es donde se ha centrado nuestra atención posteriormente, analizando la implementación de estos sistemas de recuperación por medio del estudio de su tipología (convencionales o semánticos) y determinando sus principales características en la línea de verificar la hipótesis de trabajo: la tecnología semántica se ha convertido en el principal aliado del científico para recuperar información en temas relacionados con la COVID-19. La Tabla I (ver anexo) recoge la relación de los 27 sistemas analizados, la mayoría de ellos desarrollados en Estados Unidos (13, 10 en centros de i+d y 3 en empresas), y en Europa (10, 4 en centros de i+d y 6 en empresas). La segunda de las columnas recoge información sobre la tipología de buscador implementado.

3. RESULTADOS

Se identificaron más de una cuarentena de conjuntos de datos y desarrollos de motores de búsqueda relevantes basados en CORD-19 realizados por universidades, instituciones y empresas de inteligencia artificial, dentro de este conjunto general, 27 son sistemas de recuperación de información. Destaca enormemente el trabajo del instituto AI2 que no solo ha participado activamente en el desarrollo del corpus CORD-19, lo ha alojado como una fuente de información más en el buscador semanticscholar y además ha implementado cuatro sistemas de búsqueda de tres categorías diferentes: 'Scifact' a modo de buscador convencional que emplea el modelo vectorial para calcular la similitud entre la ecuación de búsqueda (una frase del tipo 'Mass masking reduces COVID-19 transmission rates') y los resúmenes de los artículos del corpus documental; 'Scisight' herramienta de visualización de información que explora asociaciones entre cohortes de los principales investigadores médicos y las facetas (conceptos) que aparecen en el corpus, además de los vínculos existentes entre las proteínas, células, genes y enfermedades (Pahins y otros, 2019); ScispaCy, herramienta para descargar que

analiza artículos y visualiza enlaces entre los descriptores, los nombres de instituciones y dependencias entre los términos. Finalmente, ha desarrollado 'SPIKE-CORD', herramienta de procesamiento de lenguaje natural optimizado que combina búsquedas booleanas, por expresiones regulares y análisis de la estructura sintáctico-semántica de los textos permitiendo hacer coincidir los gráficos lingüísticos que subyacen en el texto (en este proyecto participa también la Universidad de Bar-Ilán de Israel). Estas tres tipologías de sistemas de búsqueda van a servir para agrupar los sistemas identificados y proceder a su presentación.

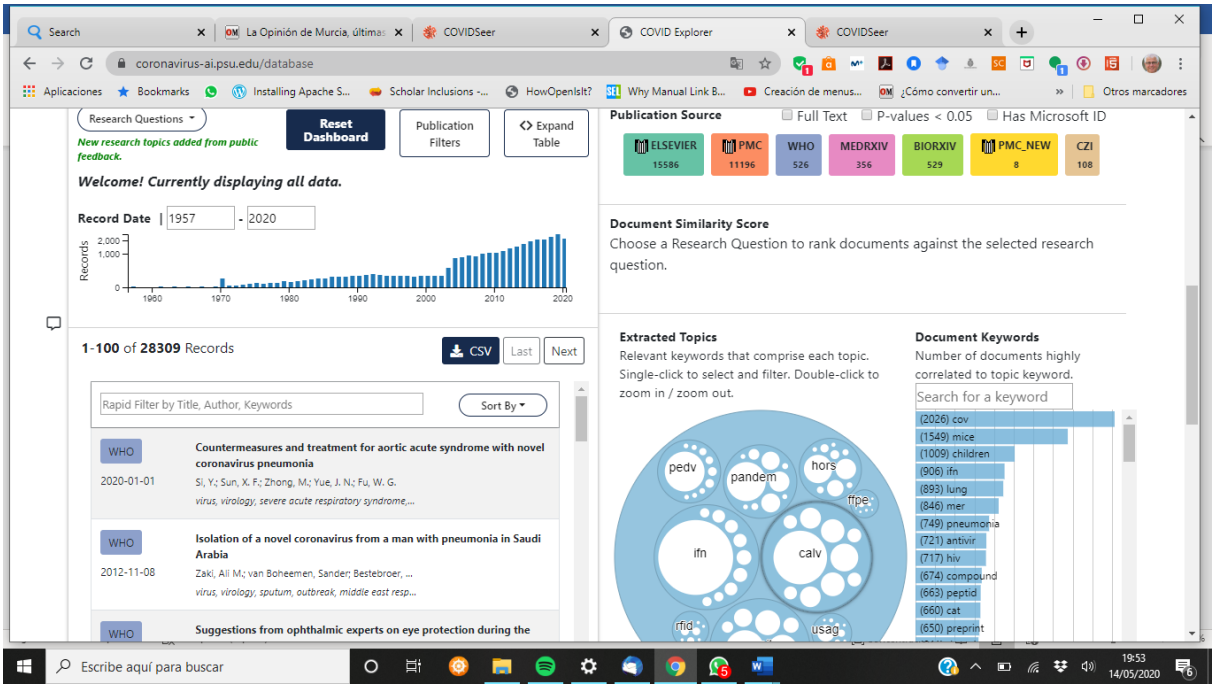
3.1. Buscadores convencionales

Además del diseñado por AI2, destaca el buscador terminológico (aunque se pueden hacer preguntas en formato de lenguaje natural) 'CORD-19 Search Vespa' que presenta los resultados con un diseño parecido al de las grandes plataformas bibliográficas y filtra documentos por repositorio fuente (PMC, Elsevier, biorxiv, WHO, etc.), revista, autor y fechas de publicación. Dotado de un diseño visual más simple, el sistema 'COVID-19 Explorer' desarrollado por el Instituto Jozëf Stefan de Eslovenia, es una herramienta que prioriza los resultados de forma interactiva llevando a cabo un algoritmo de alineamiento basado en la relevancia, si bien se apoya en un grafo de descriptores para hacer más precisas las búsquedas. Herramienta más o menos similar es la desarrollada por el Instituto de Tecnología de Gandhinagar con el sistema 'Covidexplorer' que permite búsquedas por texto libre y presenta la información en formato de nubes de etiquetas y permite refinar las búsquedas por año (desde 1957), proteínas, ADN, ARN, entidades químicas, trastornos ocasionados y tipos de células. La Universidad Estatal de Pensilvania ha desarrollado 'Covidseer', gracias a la tecnología del motor de búsqueda CiteseerX, proporcionando ayuda en la búsqueda a partir de las citas y anotaciones más destacadas del corpus documental obtenidas a partir de un proceso de extracción de información (Huang y otros, 2020). De un diseño más sencillo es la aportación de la Fundación Andrew Mellon, el buscador 'Fatcat COVID-19', prototipo de índice de búsqueda a texto completo de artículos, informes, conjuntos de datos y otros recursos de investigación relacionados con la pandemia, incluidas las respuestas rápidas del departamento de salud de Estados Unidos. El Instituto Ludwig ha creado la herramienta LIA COVID-19 ('Ludwig initiative against COVID-19') que permite localizar información por términos o por frase exacta. Este corpus en línea ha sido desarrollado para ayudar a los investigadores que no escriben en inglés de forma nativa a redactar correctamente sus comunicaciones científicas a partir del lenguaje estándar de los hablantes nativos almacenado en los miles de documentos científicos de CORD-19 (Nasution, 2018, 212). También entraría en este grupo el amplio conjunto de utilidades desarrolladas por el Ontology Engineering Group de la Universidad Politécnica de Madrid que ha desarrollado, bajo el soporte del motor de búsqueda Apache SolR, un sistema de búsqueda con grafos y estadísticas sobre la colección de artículos indexada y un explorador mediante etiquetas creadas, además de permitir llevar a cabo anotaciones de los artículos contenidos en la base de datos. Incluye también un modelo probabilístico que permite identificar los términos más descriptivos de su contenido. Además incorpora una interfaz de navegación visual lo que sitúa este sistema a medio camino entre una herramienta convencional de recuperación de información y otra perteneciente al grupo de las aplicaciones basadas en visualización de la información.

3.2. Visualización de la información

El buscador 'Covid 19 Corpus' es fruto de la implementación del motor 'Sketch Engine' sobre el corpus de documentos de CORD-19. Este sistema de búsqueda, además de la búsqueda convencional permite localizar documentos vinculados gracias a su tesauro, la frecuencia de uso, la proximidad de términos en las frases (operadores posicionales) y expresiones regulares incluidas en el corpus. Funktor ha desarrollado 'Carnap', buscador cuyos términos de búsqueda están relacionados semánticamente y se ordenan por frecuencia. Desarrollada por la Universidad de Emory, 'tmcovid' es otra herramienta de este tipo que permite extraer y resumir los bioconceptos (genes, productos químicos, fármacos, mutaciones, líneas celulares, especies y enfermedades) de la literatura científica desarrollada a causa de la pandemia COVID-19, mostrando los resultados mediante nubes de etiquetas. Microsoft ha desarrollado el buscador, 'Covid 19 Search Azure', a partir del corpus de documentos CORD-19 permitiendo la búsqueda a partir de términos semánticamente similares y filtrando los resultados por revistas, autores, trastornos ocasionados por la enfermedad, diagnósticos y tratamientos (entre otros). La Universidad Estatal de Pensilvania aporta una segunda herramienta, 'Covid Explorer' que, a partir de la aplicación de distintos algoritmos sobre el corpus documental, proporciona un cuadro de mando para la recuperación de información por fechas, fuente de información, similitud entre los documentos, nube de tópicos extraídos de los documentos y correlación del contenido de los documentos.

Figura 2. Sección del cuadro de mando del buscador Covid-Explorer



Fuente: Covid Explorer

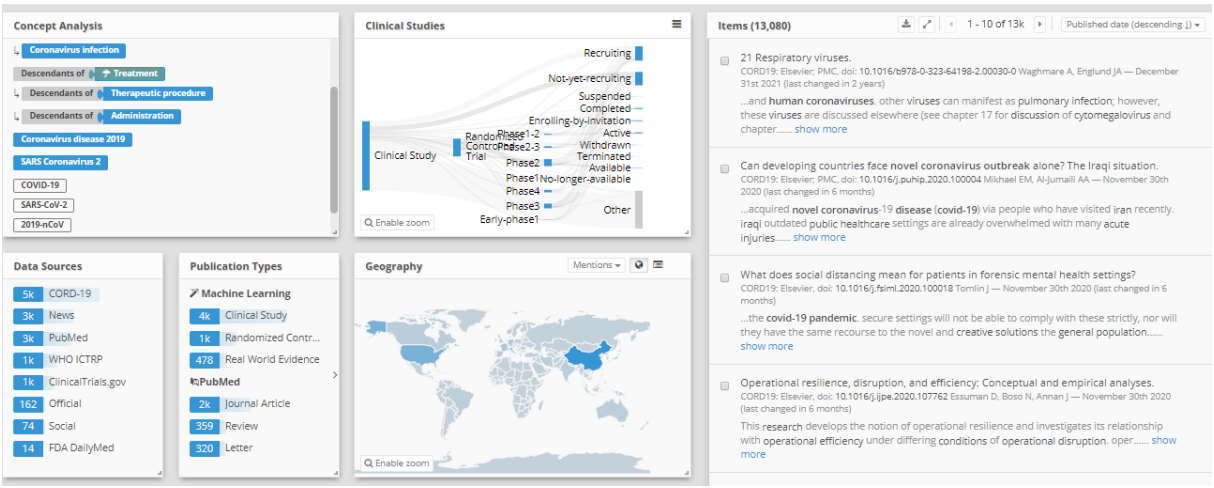
'WellAI COVID-19' es el proyecto de la empresa WellAI que crea modelos de conceptos a partir del contenido de los documentos del corpus y presenta los resultados de búsqueda en función de las asociaciones entre estos conceptos.

3.3. Proyectos de Inteligencia Artificial (PLN y extracción de términos de un corpus)

La Universidad de Ohio aporta extracción de palabras clave de la terminología 'Systematized Nomenclature of Medicine – Clinical Terms' o SNOMED CT ((Donnelly, 2006) en los artículos que forman parte del corpus por medio de la herramienta 'COVID-19 Concept Embeddings'. La Universidad de Waterloo aporta 'Neural Covidex', herramienta que aplica modelos de redes neuronales y técnicas de inteligencia artificial de última generación como parte de una suite desarrollada en pocas semanas para ayudar a los expertos en este dominio del conocimiento a enfrentarse a la pandemia reforzando la calidad de la información necesaria para la toma de decisiones basada en evidencias y en la generación de ideas (Zhang y otros, 2020). Además de la búsqueda por términos (soportada por Solr) permite la navegación por facetas gracias a modelos de redes neuronales. 'Oumy' es un buscador semántico que interacciona en lenguaje natural con los usuarios a modo de chatbot. La Universidad de Corea ha desarrollado 'CovidAsk', sistema parecido al anterior, pero que devuelve como resultado fragmentos localizados en el corpus de documentos que maneja. La Universidad de Hong Kong ha implementado 'CaiRE-Covid', sistema que, a partir de preguntas en lenguaje natural que procesa previamente a la consulta del usuario y recupera el mayor número posible de publicaciones académicas relevantes con la consulta. El segundo módulo lleva a cabo una síntesis de documentos asociados mediante red neuronal con párrafos largos que mejoran la legibilidad de los resultados generando una lista de los fragmentos de texto más relevantes de los documentos recuperados, resaltando también las palabras clave relevantes. El tercer y último módulo genera un resumen conciso de los documentos más relevantes (Su y otros, 2020). 'CORD-19 Search' es la aportación de Amazon a partir de su infraestructura de almacenamiento AWS. Ha diseñado un sistema de búsqueda en lenguaje natural cuyo resultado final es un portal de diseño similar al de las plataformas bibliográficas en el cual se puede refinar la búsqueda por tópicos y fecha de publicación. De naturaleza muy similar, pero incorporando autoaprendizaje para la mejora de las búsquedas, es la aportación de la empresa Curiosity que ha diseñado el portal 'Covid Dataset Search' que permite búsquedas terminológicas (versión experimental) y navegar por términos,

revistas, trastornos, tópicos y abreviaturas. Este buscador emplea sinónimos basados en aprendizaje automático para localizar documentos de similar contenido. Sinequa ha desarrollado otro buscador que hace uso de técnicas de inteligencia artificial denominado 'COVID-19 Intelligent Insight' que recupera información a partir de un corpus ampliado de documentos (más de 88000 artículos y/o preprints) e incorpora una amplia selección de filtros para refinar las búsquedas (colecciones, tipos de documentos, revistas, autores, lugares, etc.) y de sugerencias de temas para ampliarlas. 'Covidscholar', proyecto de la Universidad de Berkeley, también emplea procesamiento de lenguaje natural para mejorar las búsquedas en su corpus documental relacionado con la pandemia. Permite filtrar por tipo de documento, etiquetas, año y repositorio fuente. Este buscador aprovecha los modelos de aprendizaje automático que extraen el conocimiento de la literatura y ayudan a los investigadores a hacer nuevas conexiones que podrían haberse perdido debido al gran volumen de investigación que sale todos los días (Bao, 2020). El Instituto de Tecnología de Karlsruhe ha desarrollado el motor 'Discovid', herramienta que aplica un enfoque de aprendizaje automático conocido como modelado de temas que ayuda a descubrir temas subyacentes en todo el conjunto de publicaciones (bajo este enfoque, todos los artículos pueden verse como una mezcla de estos temas, pudiendo encontrarse documentos relacionados con una mezcla de temas similares). El último sistema analizado es la herramienta 'COVID-19 Doc Search Engine' de la empresa Doctor Evidence. Se trata de una muy completa herramienta que, a partir de búsqueda por conceptos y procesamiento de los textos bibliográficos por medio de procesamiento de lenguaje natura entrega un informe muy completo en la respuesta: asociaciones entre conceptos, filtro por fuentes, edad de los pacientes, género, tipo de publicaciones, fuente de los documentos, características del trastorno que sufren, tratamientos aplicados y resultados, cuáles son las palabras clave más relacionadas con el concepto, fechas de publicación y actualización, estado de evolución de los estudios, cuáles son los autores más influyentes en el campo de la consulta y otros tipos de asociaciones. Esta herramienta aporta un cuadro de mando a partir del cual el investigador puede proseguir con su tarea suficientemente informado. Es también digno de destacar el trabajo del grupo de investigación Ixa, de la Universidad del País Vasco, que han preparado un sistema de recuperación que primero filtra dentro del corpus documental los documentos más afines a la necesidad informativa, aplicando posteriormente técnicas avanzadas de recuperación de información basadas en redes neuronales de inteligencia artificial. En concreto, estas técnicas emplean el modelo lingüístico denominado BERT ('Bidirectional Encoder Representations from Transformers', utilizado en el buscador de Google) que es capaz de crear una representación contextual para cada palabra en función de las palabras que la rodean: "aquellas que tienen un significado parecido estarán más cerca entre ellas que las que no lo tienen, como si de un mapa se tratara" (Otegi y otros, 2020). Esta herramienta está disponible como un recurso más para descarga (como la herramienta Scispay del Instituto Allen) pero no está en línea.

Figura 3. Sección del cuadro de mando del buscador 'Covid-19 Doc Search Engine'



Fuente: Covid-19 Doc Search Engine

Es ciertamente muy considerable el número de empresas e instituciones que, a partir del fondo documental de CORD-19, han desarrollado servicios de consulta y han puesto estas herramientas a disposición de la comunidad científica. De la lista de analizada, solo 7 de ellos se encuadran dentro del campo de los buscadores convencionales (terminológicos) frente a 18 buscadores que emplean algún tipo de tecnología semántica o de procesamiento de lenguaje natural y se apoyan en técnicas de visualización de información. Es un hecho que la pandemia ha disparado el uso de buscadores semánticos frente a los convencionales por la necesidad de filtrar las operaciones de recuperación de información debido a tres razones fundamentales: (1) la enorme producción científica que llega a infoxicar, algo consustancial al tiempo presente según dice la profesora Eva Méndez (Salas, 2020); (2) la necesidad de recuperar por facetas o conceptos más que por coincidencia de términos entre la ecuación de búsqueda y el contenido de los artículos, los buscadores basados en el modelo vectorial adolecen de esa capacidad y (3) la imposibilidad material de emplear los índices de citas como referencia para la selección de artículos y, por consiguiente, la ausencia de información para la aplicación correcta de Pagerank o algoritmos similares de citación.

3.4. Posibilidades de los sistemas de búsqueda

Como es lógico, no todos los sistemas de búsqueda analizados poseen las mismas posibilidades de búsqueda y, especialmente, de presentación de los resultados (las interfaces de usuario van desde unas simples páginas de resultados a partir de una caja de búsqueda a sofisticados sistemas de visualización de la información).

En la Tabla II (ver Anexo) se han recogido de forma sintética las principales capacidades de búsqueda de cada uno de estos sistemas. Todos permiten la búsqueda utilizando operadores booleanos. Un tercio de los sistemas de búsqueda (9) permiten localizar documentos mediante expresiones regulares, concordancia o truncamiento. La inmensa mayoría de los buscadores (25 de 27) permiten recuperar información por búsqueda literal, por lemas extraídos del corpus documental o por descriptores de un tesauro. También es alto (20 de 27) el porcentaje de sistemas que permiten filtrar los resultados con base en distintos criterios (revista, autores, fechas, etc.), y 10 de estos sistemas permiten al usuario elegir la fuente documental sobre la que hacer la búsqueda, la mayor parte de los sistemas permiten recuperar información sobre el corpus documental CORD-19 o subconjuntos de este corpus, pero hay varios que expanden su alcance a fondos de editoriales publicados en abierto o a repositorios. Solo 3 sistemas permiten la descarga del corpus documental gestionado.

4. CONCLUSIONES

El volumen de la producción científica y de documentación técnica generada en torno a la COVID-19 ha requerido de métodos automatizados para su análisis y síntesis. La comunidad editorial e informática ha estado a la altura de las circunstancias con el desarrollo de un amplio conjunto de motores y sistemas de búsqueda diseñados para servir a los investigadores.

El éxito del conjunto de datos CORD-19 se materializa no solo en la cantidad de fuentes que lo utilizan sino también en el amplio número de buscadores terminológicos y semánticos desarrollados aprovechando su corpus gracias a su rápida disposición en alojadores y buscadores de conjuntos de datos, como es Kaggle. El soporte proporcionado por el motor 'Semantic Scholar' de AI2 y Microsoft, ha permitido una mayor difusión y explotación de la literatura de CORD-19, convirtiéndose así en el referente en las búsquedas bibliográficas. No solo Microsoft ha aportado esfuerzos al desarrollo de la investigación contra la pandemia, algunas de las principales empresas de tecnología han aparecido como auspiciadoras de estos sistemas de recuperación (Amazon y Facebook son dos excelentes botones de muestra) y, fuera del ámbito de este estudio, otras 'sitcoms' han puesto su granito de arena (IBM, la más antigua empresa informática, ha conectado los principales supercomputadores del mundo para apoyar al desarrollo de modelos epidemiológicos que permitieran conocer el desarrollo de la enfermedad).

La simbiosis entre la tecnología que permite gestionar documentos y luego recuperarlos ha sido de muy alto nivel. Si bien habría que desarrollar un estudio específico sobre usabilidad, experiencia de uso y satisfacción del usuario final de estos sistemas, aquellos que aúnan el procesamiento de lenguaje natural con una óptima visualización de los resultados parecen los candidatos ideales para que su uso se expanda entre la comunidad científica (en esto destacan especialmente los sistemas 'Covid-Explorer' de la Universidad de Pensilvania y el buscador 'Covid-19 Doc Search Engine' de la empresa DRE Evidence, además del propio buscador 'Semantic Scholar').

Además de CORD-19, el conjunto de fuentes que manejan los sistemas de recuperación de información desarrollados (los más completos) es más amplio, lo que representa también un esfuerzo de interoperabilidad semántica de datos y documentos, otro interesante avance máxime si se tiene en cuen-

ta el poco tiempo que ha pasado desde que estas herramientas comenzaron a funcionar. Queda ahora averiguar si este paso adelante de la tecnología semántica se va a quedar circunscrito a la lucha contra la pandemia o si se va a ampliar a otros sistemas de información, lo lógico es que así fuera.

5. REFERENCIAS

Adams, J., Light, R. (2020). *What Role Does Collaboration have in Responding to COVID-19?* https://osf.io/preprints/socarxiv/jqwyr/

Bao, Y., Bossion, A., Brambilla, D., Buriak, J. M., Cai, K., Chen, L., Horton, M. K. (2020). Snapshots of Life—Early Career Materials Scientists Managing in the Midst of a Pandemic. *Chemistry of Materials*, 32 (9), 3673-3677. https://doi.org/10.1021/acs.chemmater.0c01624

Baumann N. (2016). How to use the medical subject headings (MeSH). International. *Journal of Clinical Practice*, 70(2). pp.171-174. https://doi.org/10.1111/ijcp.12767

Callaghan S. (2020). COVID-19 Is a Data Science Issue. *Patterns*, 1 (2), 100022. preprint. https://doi.org/10.1016/j.patter.2020.100022

Colavizza, G., Costas, R., Traag, V. A., Van Eck, N. J., Van Leeuwen, T., Waltman, L. (2020). A scientometric overview of CORD-19. *BioRxiv*. https://doi.org/10.1101/2020.04.20.046144

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121, 279-290.

Dousset, B., Mothe, J. (2020). Getting Insights from a Large Corpus of Scientific Papers on Specialisted Comprehensive Topics--the Case of COVID-19. *arXiv preprint.* https://arxiv.org/abs/2005.00485

Fernández-Sellers, M.; Acedo J.; Lozano-Tello, A. (2019). Identification of representative terms of datasets. *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, Coimbra, Portugal, pp. 1-6.

Huang, T. H. K., Huang, C. Y., Ding, C. K. C., Hsu, Y. C., Giles, C. L. (2020). CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. *arXiv preprint.* https://arxiv.org/abs/2005.02367

Kousha, K., Thelwall, M. (2020). COVID-19 publications: Database coverage, citations, readers, tweets, news, Facebook walls, Reddit posts. *Quantitative Science Studies*, 1 (3), 1068-1091. https://doi.org/10.1162/qss_a_00066

Nasution, D. K. (2018). Corpus Based-Approach in Enhancing Students' Academic Writing Skill: Its Efficacy and Students *Perspectives. International Journal*, 6 (2), 210-217. https://doi.org/10.15640/ijll.v6n2a25

Otegi, A.; Soroa, A.; Agirre, E. y Campos, J.A. (2020). *Cómo gestionar la sobrecarga de información científica sobre COVID-19*. https://theconversation.com/como-gestionar-la-sobrecarga-de-informacion-cientifica-sobre-covid-19-138651

Pahins, C. A., Omidvar-Tehrani, B., Amer-Yahia, S., Siroux, V., Pepin, J. L., Borel, J. C., Comba, J. L. (2019). COVIZ: a system for visual formation and exploration of patient cohorts. *Proceedings of the VLDB Endowment*, 12 (12), 1822-1825. https://doi.org/10.14778/3352063.3352075

Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Hersh, W. R. (2020). TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19. *Journal of the American Medical Informatics Association*, 27 (9), 1431-1436. https://doi.org/10.1093/jamia/ocaa091

Salas, J. (2020, 5 de mayo). Sepultados bajo la mayor avalancha de estudios científicos. *El País*. https://elpais.com/ciencia/2020-05-04/sepultados-bajo-la-mayor-avalancha-de-estudios-cientificos.html

Su, D., Xu, Y., Yu, T., Siddique, F. B., Barezi, E. J., Fung, P. (2020). CAiRE-COVID: A Question Answering and Multi-Document Summarization System for COVID-19 Research. *arXiv preprint.* https://arxiv.org/abs/2005.03975

Torres-Salinas, D. (2020). Ritmo de crecimiento diario de la producción científica sobre Covid-19. Análisis en bases de datos y repositorios en acceso abierto. *El profesional de la información*, 29 (2). https://doi.org/10.3145/epi.2020.mar.15

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D. (2020a). CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint.* https://arxiv.org/abs/2004.10706

Wang, X., Liu, W., Chauhan, A., Guan, Y., Han, J. (2020b). Automatic Textual Evidence Mining in COVID-19 Literature. *arXiv preprint.* https://arxiv.org/abs/2004.12563

Zhang, E., Gupta, N., Nogueira, R., Cho, K., Lin, J. (2020). Rapidly Deploying a Neural Search Engine for the COVID-19 Open Research Dataset: Preliminary Thoughts and Lessons Learned. *arXiv preprint.* https://arxiv.org/abs/2004.05125

ANEXO

Tabla I. Sistemas de recuperación de información desarrollados a partir del corpus CORD-19 para apoyar la investigación contra la COVID-19

Recurso basado en CORD-19	Tipología	Desarrollador	URL
CaiRE-Covid	(IA)	Centre for Artificial Intelligence of the University of Hong Kong. China. **(i+d)**	https://caire.ust.hk/
Carnap	(visual)	Funktor. Turquía. **(empresa)**	https://carnap.ai
CORD-19 Search	(IA)	Amazon. Estados Unidos. **(empresa)**	https://cord19.aws
CORD-19 Search Vespa	(conv)	The Vespa Engine. **(i+d)**	https://cord19.vespa.ai
Covid 19 Corpus	(visual)	Lexical Computing Limited. Reino Unido. **(empresa)**	http://ske.li/covid_19
Covid Dataset Search	(IA)	Curiosity. Alemania. **(empresa)**	https://covid.curiosity.ai
Covid Explorer	(visual)	Pennsylvania University. Estados Unidos. **(i+d)**	https://coronavirus-ai.psu.edu/database
COVID-19 Concept Embeddings	(IA)	Ohio State University. Estados Unidos. **(i+d)**	https://slate.cse.ohio-state.edu/JET/COVID-19/
COVID-19 DOC Search Engine	(IA)	Doctor Evidence. Estados Unidos. **(empresa)**	https://covid19.drevidence.com/
COVID-19 Intelligent Insight	(IA)	Sinequa. Reino Unido. **(empresa)**	https://covidsearch.sinequa.com
Covid-19 Search Azure	(visual)	Microsoft. Estados Unidos **(empresa)**	https://covid19search.azurewebsites.net
Covid19 Explorer	(conv)	Institute Jožef Stefan. Eslovenia. **(i+d)**	http://covid19explorer.ijs.si
CovidAsk	(IA)	Korea University. Corea. **(i+d)**	https://covidask.korea.ac.kr/
Covidexplorer	(conv)	Indian Institute of Technology. India. **(i+d)**	http://covidexplorer.in
CovidScholar (Matscholar)	(IA)	Lawrence Berkeley National Laboratory. Estados Unidos. **(i+d)**	https://www.covidscholar.com
CovidSeer	(conv)	Pennsylvania University. Estados Unidos. **(i+d)**	http://covidseer.ist.psu.edu
Discovid	(IA)	Karlsruhe Institute of technology. Alemania. **(i+d)**	https://discovid.ai
Fatcat COVID-19 Paper Search	(conv)	Andrew Mellon Foundation. Estados Unidos. **(i+d)**	https://covid19.fatcat.wiki
IA & COVID-19	(conv)	Polytechnic University of Madrid. Ontology Engineering Group.	https://oeg-upm.github.io/covid19/servicios/
Ixa	(IA)	Universidad del País Vasco **(i+d)**	http://ixa2.si.ehu.es/convai/kaggle-cord19/round1.html
LIA COVID-19	(conv)	Ludwig. Italia. **(empresa)**	https://covid19.ludwig.guru
Neural Covidex	(IA)	University of Waterloo. Canadá. **(i+d)**	https://covidex.ai
SciFact	(conv)	The Allen Institute for Artificial Intelligence. Estados Unidos. **(i+d)**	https://spike.covid-19.apps.allenai.org/search/covid19
Scisight	(visual)	The Allen Institute for Artificial Intelligence. Estados Unidos. **(i+d)**	https://scisight.apps.allenai.org/
ScispaCy	(visual)	The Allen Institute for Artificial Intelligence. Estados Unidos. **(i+d)**	https://allenai.github.io/scispacy
tmCOVID	(visual)	Emory University. Estados Unidos. **(i+d)**	http://tmcovid.com
WellAI COVID-19	(visual)	Wella. Alemania. **(empresa)**	https://wellai.health/covid

Leyendas: **conv**: buscador convencional; **(visual)** visualización de la información; **(IA)** inteligencia artificial; **(empresa)** organización con ánimo de lucro **(i+d)** organización dedicada a la investigación.

Tabla II. Posibilidades de búsqueda en los SRI analizados

Recurso basado en CORD-19	Tipo	Booleana	Expresiones regulares	Exacta	Filtrado	Fuente	Descarga
CORD-19 Search Vespa	(conv)	X		X	X	X	X
Covid19 Explorer	(conv)	X		X			X
Covidexplorer	(conv)	X			X		
CovidSeer	(conv)	X		X	X	X	
Fatcat COVID-19 Paper Search	(conv)	X	X	X	X	X	
IA & COVID-19	(conv)	X		X		X	
LIA COVID-19	(conv)	X	X	X	X		
SciFact	(conv)	X	X		X		
CaiRE-Covid	(IA)	X	X	X			
CORD-19 Search	(IA)	X		X			
Covid Dataset Search	(IA)	X		X	X		X
COVID-19 Concept Embeddings	(IA)	X		X			
COVID-19 DOC Search Engine	(IA)	X		X	X		
COVID-19 Intelligent Insight	(IA)	X	X	X	X	X	
CovidAsk	(IA)	X		X			
CovidScholar (Matscholar)	(IA)	X		X	X	X	
Discovid	(IA)	X		X	X		
Ixa	(IA)	X		X			
Neural Covidex	(IA)	X	X	X	X	X	
ScispaCy	(IA)	X	X	X			
Carnap	(visual)	X		X	X		
Covid 19 Corpus	(visual)	X	X	X	X		
Covid Explorer	(visual)	X		X	X	X	
Covid-19 Search Azure	(visual)	X	X	X	X		
Scisight	(visual)	X		X	X	X	
tmCOVID	(visual)	X		X	X	X	
WellAI COVID-19	(visual)	X		X	X		

Leyendas: **conv**: buscador convencional; **(visual)** visualización de la información; **(IA)** inteligencia artificial.