La investigación del SARS-CoV2 mediante el uso de datos abiertos y grafos de conocimiento

Eder Ávila Barrientos
Instituto de Investigaciones Bibliotecológicas y de la Información de la UNAM, México

Introducción

De acuerdo con la Organización Mundial de la Salud (OMS),[1] los coronavirus son una extensa familia de virus que pueden causar enfermedades tanto en animales como en humanos. En los humanos, se sabe que varios coronavirus causan infecciones respiratorias que pueden ir desde el resfriado común hasta enfermedades más graves como el síndrome respiratorio de Oriente Medio (MERS) y el síndrome respiratorio agudo severo (SRAS). El coronavirus que se ha descubierto más recientemente causa la enfermedad por coronavirus COVID-19.

En este contexto, la investigación científica, médica y académica tienen dos propósitos fundamentales, el primero de ellos está relacionado con el estudio del comportamiento del virus SARS-CoV2 que permita comprender su transmisión entre los seres humanos y animales. Por otro lado, conocer la estructura genética de la

1 Organización Mundial de la Salud, "Preguntas y respuestas sobre la enfermedad por coronavirus (COVID-19)", https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses.

enfermedad COVID-19 para generar una vacuna o antídoto que permita erradicar su contagio y letalidad. En ambos casos, los datos son elementos esenciales, pues sin ellos todo proceso de investigación quedaría reducido a meros esfuerzos sin resultados exitosos.

Los datos generados como parte de esta pandemia son de diversa naturaleza y tipología. Por un lado, se observan datos estadísticos y numéricos que son utilizados para representar la cantidad de decesos, contagios y estimaciones futuras del impacto de la COVID-19 en los seres humanos. La mayoría de estos datos han sido liberados abiertamente para su consulta y reutilización, lo cual ha propiciado el desarrollo de diversas aplicaciones que emplean representaciones para explicar el comportamiento de la enfermedad mediante el análisis y procesamiento de los datos. Por otra parte, en diversas plataformas digitales se puede acceder a datos de índole científica que tienen el propósito de estudiar el comportamiento del virus SARS-CoV2; sobre todo se hace referencia a datos genéticos, bioquímicos y clínicos. Con estos datos se pretende estudiar la estructura del virus y descifrar su comportamiento al momento de transmitirse de persona a persona o entre mamíferos.

Figura 1. Coronavirus Resource Center

Fuente: Johns Hopkins University, 2020.
Disponible en https://coronavirus.jhu.edu/map.html.

En la figura 1, puede apreciarse una de las principales fuentes de datos que fueron generadas como parte del estudio y análisis del progreso de la pandemia. Se trata de un centro de datos desarrollado por la Johns Hopkins University con base en los datos oficiales de la OMS. El uso de datos abiertos ha potencializado la generación de este tipo de plataformas, lo que ha permitido obtener desarrollos significativos y reales que ponen en evidencia la importancia de contar con datos en cantidades masivas, pero procesados de manera integral para obtener diversas representaciones, lo cual permite realizar análisis de distintas variables que intervienen en el fenómeno de la pandemia.

Bajo esta premisa, los grafos del conocimiento son representaciones complejas que permiten visualizar y acceder a datos, recursos y contenidos en un escenario común, el cual es conocido como dominio de especificación. Estos grafos tienen la característica principal de ofrecer un método intuitivo para descubrir piezas de conocimiento y las respectivas vinculaciones que se establecen en el dominio al que pertenecen.

Se les denomina 'grafos de conocimiento' por su amplia utilidad en la identificación de patrones ocultos, significativos e interpretativos en amplios dominios de datos relacionados con recursos y contenidos que forman parte de una determinada problemática o fenómeno. La fuente principal de estos grafos son los datos. Se observa que en tiempos de pandemia, diversos sistemas y plataformas han optado por la utilización de estas representaciones visuales debido a la alta complejidad que el análisis y la interpretación de la pandemia ha manifestado.

De esta manera, el objetivo de este trabajo consiste en analizar el uso de los datos abiertos y los grafos de conocimiento en la investigación del SARS-CoV2, pues se trata de un binomio que ha hecho una fuerte presencia en el ámbito indagatorio y analítico de la pandemia y en una amplia gama de sus vertientes.

La metodología para la conformación de este trabajo está conformada por dos fases. En la primera de ellas, se ha realizado un proceso de revisión de la literatura especializada. Mediante la hermenéutica y el análisis del discurso aplicados a esta literatura,

se han identificado una serie de hallazgos que permiten estimar un mayor avance en el estudio y la comprensión del nuevo coronavirus mediante el uso de fuentes de datos libres de restricciones económicas, técnicas y legales. Además, se han identificado conjuntos de datos liberados de manera abierta para su reutilización en la investigación de SARS-CoV2. La segunda fase ha consistido en la selección y procesamiento de un conjunto de datos relativos a la COVID-19, el cual fue publicado y liberado por la Organización Mundial de la Salud. Con este conjunto, se ha desarrollado una ejemplificación del uso de grafos de conocimiento para el análisis de las variables de la enfermedad a través de estrategias de recuperación y descubrimiento de datos.

El trabajo se encuentra dividido en tres apartados específicos. En el primero de ellos, denominado "investigar en tiempos de pandemia", se reflexiona acerca de los procesos y métodos actuales que intervienen en la indagación y búsqueda de resultados, pues la investigación en estos tiempos se ha visto determinada por el factor tecnológico y digital que impera en el uso de diversos tipos de herramientas.

A su vez, en el segundo apartado, titulado "los datos abiertos y su aplicación en la investigación del SARS-CoV2", se aborda una serie de consideraciones teóricas respecto al uso de fuentes de datos para el descubrimiento de nuevos hallazgos relacionados con el estudio del SARS-CoV2, pues se observa que el uso de estas fuentes ha contribuido notablemente al desarrollo de aplicaciones que aceleran el proceso de descubrimiento e intercambio de datos en el contexto de la investigación científica y académica.

Finalmente, en el tercer apartado, intitulado "los grafos de conocimiento y su uso en la visualización de datos abiertos de la pandemia", se analiza el papel de la representación y visualización de los datos mediante métodos interactivos e intuitivos que son aplicados en aplicaciones y plataformas a través del uso de grafos de conocimiento. Se estima que el uso de estos grafos propicie una mayor compresión de las variables que caracterizan a la pandemia y permita descubrir patrones que nos son perceptibles e identificables por un sistema de búsqueda y recuperación de

información convencional, pues el resultado del procesamiento de datos a gran escala se resume en la generación de métodos visuales que permitan descubrir nuevos conocimientos con base al uso y explotación de datos altamente significativos.

Investigar en tiempos de pandemia

En la historia de la humanidad han acontecido pandemias que han marcado el rumbo de las civilizaciones. La lucha contra los virus y las bacterias ha sido una constante que ha contribuido al avance científico y su aplicación en la cura de enfermedades principalmente a través de la generación de medicamentos y vacunas. Las pandemias, desde la óptica indagadora, son una problemática con diversas aristas que versan desde lo social y científico, hasta lo económico y cultural.

En tiempos de coronavirus, será trascendental contar con los mecanismos idóneos para contener los estragos de la pandemia. Esto se traduce en efectuar investigaciones de mayor alcance en cortos periodos de tiempo, pues los virus son agentes que mutan constantemente y su comportamiento dinámico libera enormes cantidades de datos. Los datos son aquellos elementos que ayudan a determinar el comportamiento de los virus, además su procesamiento es crucial para identificar múltiples variables de su naturaleza. Por ejemplo, estructura genética, niveles de contagio y estragos en los índices económicos y socioculturales de la población. Bajo esta premisa, "la tasa de contagio y la letalidad han sido muy diferentes entre los países. Estas diferencias, responden, por un lado, a la respuesta de cada país ante la pandemia. Pero por otro, a patrones de poblaciones diferentes y a la variabilidad en la forma de aportar los datos".[2]

2 Mira, J. "Pandemia COVID-19: y ahora ¿qué?", *Journal of Healthcare Quality Research* 785 (2020), https://doi.org/10.1016/j.jhqr.2020.04.001.

Para enfrentar las grandes problemáticas que esta pandemia ha generado en la actualidad, la investigación se ha concebido como una actividad colaborativa apoyada notablemente por el uso de las tecnologías de la información y comunicación, sobre todo aquellas de índole computacional. De hecho, el procesamiento intensivo de datos y su respectivo descubrimiento son dos cuestiones inherentes al desarrollo de la tecnología computacional, pues la rapidez e inmediatez del tratamiento de los datos es una constante que moldea esta realidad datificada. Al respecto, Mundie[3] manifiesta que "estamos acumulando cantidades de datos en forma digital que anteriormente eran inimaginables, datos que contribuirán a desencadenar una profunda transformación en la investigación y la comprensión científica".

Por lo tanto, el procesamiento de datos tiene el gran reto de descifrar la respuesta a los problemas mediante el análisis de los datos, acelerar la comprensión de la pandemia para identificar el comportamiento y la estructura del SARS-CoV2 para paulatinamente desarrollar su antídoto. La investigación en tiempos de pandemia, además de ser colaborativa, requiere de un amplio sentido humanístico y social, contemplar que en diferentes contextos de la sociedad, los efectos de la propagación del virus han sido muy diversos, pues

> […] ha sido más audaz, y su desfachatez nos ha revelado algo que ya sabíamos, pero no lográbamos calibrar del todo: la pluralidad de niveles en que estamos conectados los unos a los otros, así como la complejidad del mundo que habitamos, de sus dinámicas sociales, políticas, económicas e incluso interpersonales y psíquicas.[4]

Sin la presencia y recolección de los datos, la investigación quedaría reducida a mera superstición, sin fundamento ni materia prima

---

3 Craig Mundie, "El camino por recorrer". En *El cuarto paradigma: descubrimiento científico intensivo de datos*, ed. Tony Hey, Stewart Tansley y Kristin Tolle. (México: Universidad Autónoma Metropolitana, 2009), 241.

4 Paolo Giordano, *En tiempos de contagio*. (España: Salamandra, 2020), 1.

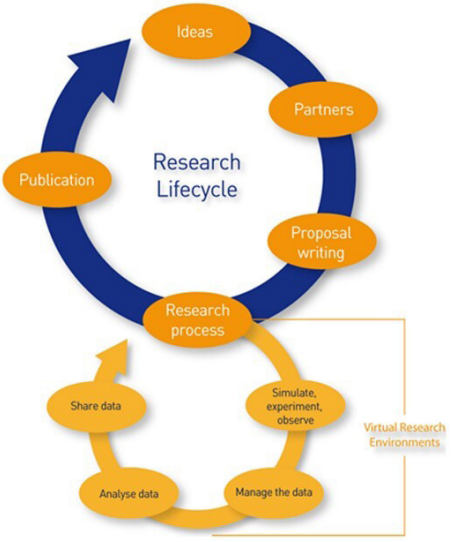
para el desarrollo de experimentaciones, indagaciones y reflexiones. La realidad se encuentra cimentada en datos, pues éstos la representan de diferente forma y con muy variada naturaleza. Así, la colaboración científica en la actualidad es desarrollada de manera virtual mediante el uso de grandes cantidades de datos y la explotación de las tecnologías computacionales que permiten la comunicación sincrónica y asincrónica de manera remota. Nuevas metodologías para la recolección de datos han sido adaptadas por diversos sectores de investigación. Además, la multidisciplinariedad de la investigación es una constante que ha permitido establecer grupos de investigación con especialistas de diversos campos del conocimiento. Aunado a ello, las fuentes de datos abiertos han sido trascendentales para el estudio del virus y la enfermedad que ha desencadenado.

Los datos abiertos y su aplicación en la investigación del SARS-CoV2

Las redes de colaboración científica han transformado y modificado su accionar y la manera de desarrollar investigación al adaptar sus alcances en la aplicación de nuevas metodologías para el estudio del SARS-CoV2 de manera integral pero también virtual. Al respecto, los entornos virtuales de investigación han sido implementados de una manera vertiginosa como respuesta a la demanda que la pandemia ha provocado, sobre todo respecto a la aceleración en la obtención de nuevos conocimientos que permitan hacer frente al coronavirus.

Si bien este tipo de entornos ya eran utilizados antes de la pandemia, en la actualidad su utilización resulta trascendental para la colaboración y el intercambio de resultados indagatorios de una manera remota y en respuesta a las características del propio fenómeno. Bajo este contexto, conviene abordar la interacción entre la investigación y el ciclo de vida de los datos en un ambiente caracterizado por el uso de las tecnologías digitales.

Figura 2. Etapas de la investigación y el ciclo de vida de los datos



Fuente: https://doi.org/10.1371/journal.pone.0021101.g001.

Para propiciar la reutilización de datos es necesario ejercer buenas prácticas y sujetarse a los principios de su ciclo de vida para generarlos, recopilarlos, administrarlos, analizarlos y compartirlos.[5] Las etapas de la investigación y el ciclo de vida de los datos guardan una estrecha relación. Cuando ambos elementos son trasladados a un ambiente digital, entonces se configura un marco centralizado en su utilización en el proceso indagatorio, ya sea para la validación o refutación de una hipótesis o bien, para la obtención de resultados. En la figura 2 puede apreciarse la manera de interactuar entre el ciclo de vida de los datos, el proceso de investigación y los entornos virtuales de investigación.

5 Carol Tenopir *et al.*, "Data Sharing by Scientists: Practices and Perceptions", *PLoS ONE,* núm. 6 (2011), https://doi.org/10.1371/journal.pone.0021101, 2.

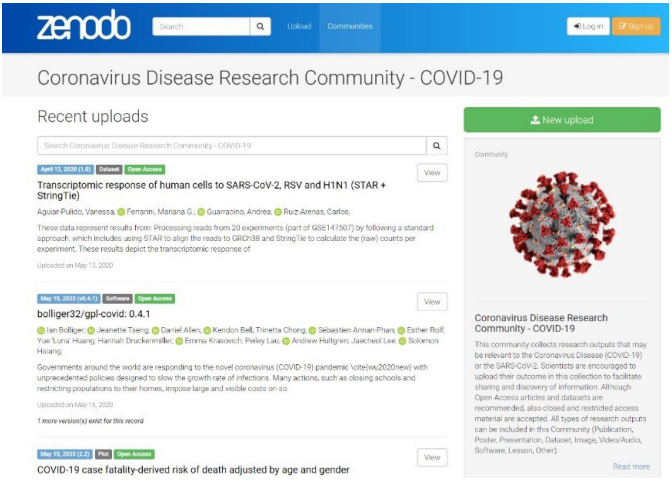
El foco central de esta interacción son los datos *per se*, pues son la materia prima en donde se establecen los elementos para la fundamentación, experimentación y obtención de resultados. Mediante simulaciones, experimentos y observaciones, se obtienen datos que son gestionados mediante metodologías y herramientas; a su vez, el procesamiento de los datos permite su posterior análisis para su latente compartición en un proceso de accesibilidad motivado por el factor de su reutilización. Además, sin el papel de los datos sería imposible establecer relaciones lógicas entre los objetos de estudio y los problemas que se desea indagar como parte de una problematización. Aunado a ello, los datos se constituyen a partir de la adopción de diferentes sistemas de prácticas en las comunidades científicas, los cuales son desarrollados a través del tiempo y actualmente mediante la conformación de plataformas digitales.

Por ejemplo, los repositorios de datos como Zenodo (https://zenodo.org/) han desarrollado comunidades digitales en donde es posible acceder a conjuntos de datos relacionados con tópicos multidisciplinarios vinculados con el desarrollo de la investigación acerca de la pandemia, ya que "recopila resultados de investigación que pueden ser relevantes para la enfermedad del coronavirus (COVID-19) o el SARS-CoV2. Además de alentar a los científicos a subir sus resultados en dicha colección para facilitar el intercambio y el descubrimiento de información".[6] En la figura 3, puede apreciarse un ejemplo de comunidad creada en Zenodo referente a recursos de información, conjuntos de datos y herramientas computacionales que pueden aplicarse en la identificación de hallazgos referentes al COVID-19 y al SARS-CoV2. Resulta interesante la postura multidisciplinaria de las comunidades científicas como un método para unificar esfuerzos y publicar resultados que contribuyan a una mejor comprensión de la cadena de fenómenos y problemáticas que ha desencadenado la pandemia.

---

6 Zenodo, "Coronavirus Disease Research Community-COVID-19", https://zenodo.org/communities/covid-19/?page=1&size=20.

Figura 3. Coronavirus disease researh community, COVID-19



Fuente: https://zenodo.org/communities/covid-19/?page=1&size=20.

Al respecto, la investigación médica necesita de datos confiables para identificar patrones que les permitan comprender el comportamiento del COVID-19, una enfermedad de rápida propagación e infección que ha cobrado más de 290 mil muertes alrededor del mundo.[7] El amplio dinamismo de esta enfermedad ha provocado serios daños en los sistemas de salud. Algunos de ellos han colapsado por su incapacidad para hacer frente al amplio número de personas infectadas, lo que ha provocado serios problemas económicos y una drástica pérdida de empleos. Esta situación ha obligado a los gobiernos de todo el mundo a implementar políticas de estado para contener los efectos de la pandemia.

Diversos países alrededor del mundo y organismos internacionales han optado por liberar cantidades considerables de datos estadísticos y cuantificables de los efectos de la pandemia, liberándolos

7 Estas cifras fueron recabadas al momento de desarrollar este trabajo con base en los datos estadísticos obtenidos del COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) en Johns Hopkins University (JHU). Disponible en https://coronavirus.jhu.edu/map.html.

de restricciones económicas, técnicas y legales. Además, han hecho énfasis en la reutilización de estos datos para obtener avances en el desarrollo de las investigaciones. La apertura y disponibilidad de estos datos se encuentra en plataformas digitales que fomentan su recuperación mediante estrategias de búsqueda y acceso. En la siguiente tabla pueden apreciarse algunos ejemplos de estas plataformas:

Tabla 1. **Plataformas de datos abiertos referentes a la pandemia**

Nombre de plataforma	Tipo de datos	URL de acceso
Datos.Gob.Mx | Bases de datos COVID-19[8]	Datos estadísticos obtenidos a partir del estudio epidemiológico de caso sospechoso de enfermedad respiratoria viral al momento que se identifica en las unidades médicas del Sector Salud.	https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico
WHO Coronavirus Disease (COVID-19) Dashboard	Plataforma oficial de la Organización Mundial de la Salud que incluye datos estadísticos de las infecciones y muertes por COVID-19 alrededor del mundo.	https://covid19.who.int/?gclid=Cj0KCQjw-_j1BRDkARIsAJcfmTEzLYqd8oZv46-4otstSn83ykr2rTfXkfYACkeRJ2Ar5yMc3D7K-PU4aAr-lEALw_wcB
Humanitarian Data Exchange | Novel Coronavirus (COVID-19) Cases Data	Los datos de esta plataforma son recopilados por el Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE), la OMS y diversas entidades gubernamentales a nivel internacional.	https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases
CORD-19	Proyecto desarrollado por Semantic Scholar Group del Instituto Allen. Proporciona a los investigadores herramientas abiertas y conjuntos de datos para motivar la identificación de hallazgos sobre el nuevo coronavirus.	https://www.semanticscholar.org/cord19/get-started
EU Open Data Portal | COVID-19 Data	Plataforma oficial de la Unión Europea que permite al público en general reutilizar datos de diversa índole temática. Recientemente han desarrollado una sección enfocada al reporte de casos de COVID-19 en todo el mundo.	https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data/resource/55e8f966-d5c8-438e-85bc-c7a-5a26f4863

Fuente: elaboración propia, 2020.

8 La base de datos se encuentra codificada en formato *.csv*, lo cual facilita su descarga disponibilidad y reutilización. Además, se mantiene actualizada de acuerdo con los datos registrados en la Secretaría de Salud de México.

La apertura de los datos ha crecido enormemente en la última década. "Cada vez más conjuntos de datos se han abierto al público, las interfaces de programación de aplicaciones (APIs) han sido diseñadas para permitir al público hacer uso de datos en tiempo real y se han desarrollado nuevas aplicaciones basadas en estos datos".[9] En tiempos de pandemia, la generación de aplicaciones ha contribuido a la generación de plataformas digitales que fomentan el acceso y la consulta de los datos en una amplia variedad de formatos.

En la actualidad y bajo las circunstancias que el mundo atraviesa, los datos abiertos sirven para reutilizarlos en procesos de investigación, o bien, en la generación de métodos de representación gráfica que ayuden a mejorar la comprensión de la problemática ocasionada por el nuevo coronavirus. "La visualización a través de cuadros, gráficos e imágenes es una forma efectiva y eficiente de interpretar y comprender datos y ayudar a detectar información valiosa como patrones, tendencias y anomalías".[10] Uno de los principales productos derivados del procesamiento de datos es la generación de métodos visuales para su comprensión e identificación de patrones de comportamiento en ellos.

En este sentido, algunos datos de la pandemia han sido recabados mediante diferentes técnicas y metodologías para ser depositados en plataformas para su posterior reutilización. Por ejemplo, datos estadísticos referentes a decesos, cifras de infectados, pacientes recuperados, casos sospechosos y datos socioeconómicos derivados del impacto de la pandemia en las diferentes naciones. En la tabla 1 se muestran algunos de estos ejemplos, los cuales no son exhaustivos pero permiten obtener una visión generalizada del uso de datos abiertos en el proceso de indagación y la búsqueda de una mejor comprensión de los efectos de la pandemia.

Cabe señalar que la apertura de los datos no asegura su total reutilización, ya que para ello será necesario contar con una serie

---

9 Yannis, Charalabidis *et al., The World of Open Data Concepts, Methods, Tools and Experiences*. (Suiza: Springer, 2018), 1.

10 Nitin Kale y Nancy Jones, *Practical Analytics*. (Epistemy Press, 2020), 49.

de requisitos y habilidades que permitan manejarlos, procesarlos y aplicarlos con un propósito en específico. La reutilización de los datos conlleva el dominio de métodos, estándares, infraestructuras y herramientas para mejorar la capacidad de descubrimiento, la interoperabilidad semántica, la trazabilidad y el intercambio de datos.

De ahí la necesidad de contar con profesionales de la información especializados en el manejo y el procesamiento de los datos, pues "es casi una visión universal entre los científicos y aquellos que financian la investigación básica que la ciencia debe ser más colaborativa si se quieren lograr avances científicos futuros. Un enfoque es facilitar la gestión, el intercambio y la reutilización de datos a gran escala y a largo plazo".[11]

De esta manera, la plena identificación de las fuentes de los datos es la etapa inicial que permite establecer un proceso para su gestión y accesibilidad. La pandemia actual plantea preguntas importantes sobre la apertura, el intercambio y el uso de datos y destaca los desafíos asociados a su confiabilidad y aceptación. Estos aspectos son sumamente relevantes para la aplicación de los datos en procesos de indagación y experimentación, sobre todo cuando la búsqueda de resultados confiables es una constante.

Además, el intercambio de datos ocurre cuando los científicos intencionalmente ponen sus propios datos a disposición de otras personas para su uso en investigaciones u otros esfuerzos científicos relacionados. En este sentido, "los científicos comparten datos incluidos en sus *datasets* y en sus artículos publicados, además publican datos en sitios web institucionales o personales, depositan conjuntos de datos en repositorios o envían datos en respuesta a solicitudes personales de colegas investigadores".[12] A través de

---

11 Ixchel Faniel y Trond Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues Data", Computer Supported Cooperative Work 19, (2010), 355 https://doi-org.pbidi.unam.mx:2443/10.1007/s10606-010-9117-8 (Consultado el 20 de mayo de 2020).

12 Jilian Wallis *et al*., "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology", *Plos One (2013),* https://doi.org/10.1371/journal.pone.0067332.

la historia, el intercambio de datos ha sido una constante, solo que en tiempos actuales las tecnologías computacionales y el auge de Internet, así como los alcances de la web han acelerado el intercambio.

Por ejemplo, el National Center for Biotecnology Information (NCBI) ha desarrollado un espacio digital (disponible en https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/) en el cual es posible consultar secuencias de datos referentes al genoma del SARS-CoV2. La tabulación de los datos permite obtener una visualización de los patrones de comportamiento del virus, además de obtener una consulta detallada que les permite a los especialistas establecer principios para explicar y determinar cuál es la estructura genómica del virus.

El desconocimiento de la estructura genómica del virus ha provocado lagunas de conocimiento respecto a su comportamiento. Esta situación ha influido notablemente en el número de contagios y en el desarrollo de vacunas y antibióticos que eviten un incremento mayor en las cifras de decesos ocasionados por la enfermedad. "La necesidad de desarrollar rápidamente una vacuna contra el SARS-CoV2 llega en un momento de explosión en la comprensión científica básica, incluso en áreas como la genómica y la biología estructural, que está apoyando una nueva era en el desarrollo de vacunas".[13] La revolución de los datos en el área de las ciencias de la salud ha provocado un cambio vertiginoso en la manera de desarrollar investigación, pues a mayor cantidad de datos, mayor certeza tendrá que existir en la capacidad de procesamiento que los algoritmos tengan que ejercer para conocer en mejor grado las características del virus.

Sin embargo, a pesar de los grandes avances de la inteligencia artificial y su aplicación en la tecnología computacional, científicos e investigadores continúan buscando la respuesta a una pregunta elemental: ¿Cómo la inteligencia artificial puede ayudar a

13 Lurie Nicole, Melanie Saville, Richard Hatchett y Jane Halton. "Developing COVID-19 Vaccines at Pandemic Speed". *New England Journal of Medicine* 382, núm. 21 (2020), https://doi.org/10.1056/NEJMp2005630.

la interpretación significativa de los datos científicos? Se estima que una interpretación de estas características ayudaría a encontrar una respuesta casi inmediata a las serias problemáticas como la que el mundo adolece en la actualidad. A mayor cantidad de datos, será relevante seleccionar aquellos que ayuden a obtener un mayor conocimiento de los fenómenos; es decir, contar con datos significativos para ser procesados de manera interpretativa y establecer relaciones lógicas entre la causa, el efecto y la respuesta que caracterizan al problema.

No es de extrañar que durante la historia de la humanidad han acontecido epidemias que han diezmado a la población, muchos gobiernos y autoridades no se percataban del impacto y la letalidad de las epidemias en las personas, pues las serias dificultades para capturar y procesar a los datos causaban estragos mayores en los efectos que una determinada pandemia tenía en la sociedad. De modo que en la actualidad, además de registrar y procesar datos de una manera intensiva y extremadamente rápida, es necesario contar con mecanismos de visualización que permitan tener un acercamiento interactivo con las características del fenómeno que está sujeto al análisis. Por lo tanto, la visualización de los datos es una estrategia metodológica y procedimental que favorece la consulta de patrones que a simple vista son muy difíciles de identificar.

Los grafos de conocimiento y su uso en la visualización de los datos abiertos de la pandemia

Los grafos de conocimiento son representaciones que permiten obtener una visión integral de los datos que conforman a un contexto o fenómeno en específico. Se estima que el uso de los grafos puede ayudar a una mejor comprensión de las características de un fenómeno, esto mediante las vinculaciones de significado que se establecen entre los datos que emplean. Es precisamente el significado de los datos el elemento que otorga la posibilidad de construir estructuras de conocimiento basadas en datos reales y

altamente interpretables. De ahí que uno de los grandes retos de la inteligencia artificial sea el procesamiento interpretativo apegado al significado del mundo real al que pertenecen los datos.

Aunado a ello, en años recientes los grafos de conocimiento han emergido como un área de interés para el campo de la inteligencia artificial; sin embargo, su aplicación en la investigación básica y aplicada ha sido abordada con anterioridad por disciplinas como las matemáticas, las ciencias de la salud y ciencias computacionales.

De acuerdo con Kejriwal,[14] los grafos de conocimiento se han convertido en una representación de datos popular que se encuentra en la intersección del descubrimiento de conocimiento, la minería de datos, la web semántica y el procesamiento del lenguaje natural. Muchas aplicaciones disciplinarias utilizan estos elementos como metodologías para identificar con mayor profundidad las características de un fenómeno; sin embargo, el grafo de conocimiento integra principios teóricos y empíricos para explicar el comportamiento de los datos y su respectiva interpretación, lo cual brinde la posibilidad de descubrir nuevos hallazgos o piezas de conocimiento.

Los grafos de conocimiento a menudo se diferencian en función de su arquitectura, fines operativos, fuentes de datos, cobertura y las tecnologías utilizadas en su construcción. "Son una pieza clave para el futuro de los sistemas de inteligencia artificial y muchas otras aplicaciones que consumen y razonan con datos estructurados, incluidos motores de búsqueda, sistemas empresariales y sistemas de recomendación".[15] Además de eso, los grafos de conocimiento fomentan la organización de datos no estructurados mediante el establecimiento de estructuras flexibles para

---

14 Mayank Kejriwal, "What Is a Knowledge Graph?", en *Domain-Specific Knowledge Graph Construction*. (Suiza: Springer International Publishing, 2019), 7.

15 Musa Aliyu y Adegboyega Ojo, "Towards Building a Knowledge Graph with Open Data – A Roadmap", en International Conference on e-Infrastructure and e-Services for Developing Countries. (Lagos: EAI, 2017), 157-162.

representar a datos disponibles en diferentes fuentes, sistemas, dispositivos o plataformas; todo ello mediante el establecimiento del principio de interoperabilidad global, el cual es implementado en estándares y normas.

Tal y como sucede con las aplicaciones móviles, en la actualidad es posible conectar sistemas, dispositivos, usuarios y datos con múltiples atributos y propiedades, pues sin la interoperabilidad todos estos elementos quedarían aislados sin una comunicación entre ellos. No obstante, establecer este principio es un asunto complejo, sobre todo por la amplia heterogeneidad del entorno digital. En este sentido, es preciso recordar que una amplia gama de usuarios ha suministrado al entorno digital con una cantidad abismal de datos de muy diversa tipología. En el contexto de la apertura de los datos

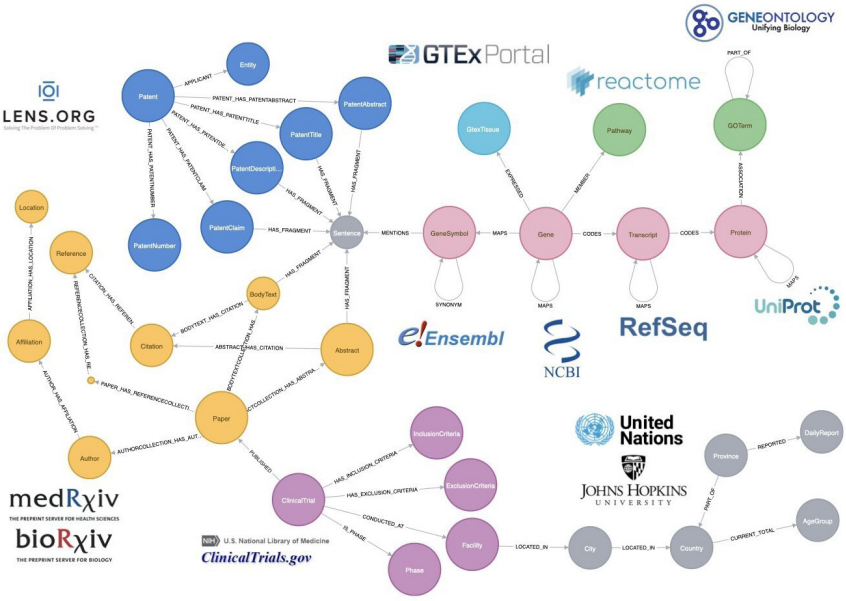
> […] la interoperabilidad de los datos abiertos es imprescindible para impulsar el movimiento de los datos abiertos vinculados y, por lo tanto, para aumentar no solo el nivel de descubrimiento y accesibilidad de los datos, sino también la posibilidad de fusionar los datos para crear nuevos escenarios de aplicación. Estos escenarios de aplicación pueden abarcar a varias partes interesadas de manera transdisciplinaria, incluidas empresas, academia, administraciones públicas y ciudadanos por igual.[16]

Por ejemplo, en la figura 4 puede observarse una representación de los datos incluidos en diferentes fuentes referentes al COVID-19, la cual es obtenida al consultar el proyecto CovidGraph desarrollado por un grupo multidisciplinario de investigación en el cual participan entidades académicas y de investigación aplicada como la Aarhus University, el German Center for Diabetes Research y la Freiburg University.

---

16 Yannis Charalabidis *et al.,* "Open Data Interoperability", en The World of Open Data: Public Administration and Information Technology, vol 28. (Suiza: Springer, 2018), 93.

Figura 4. CovidGraph representation



Fuente: https://live.yworks.com/covidgraph/.

CovidGraph contiene datos abiertos acerca de publicaciones, patentes, estructuras genéticas y conjuntos de datos relacionados con enfermedades generadas a partir de la proliferación del COVID-19. En este proyecto, se hace uso de grafos de conocimiento para identificar patrones de interacción entre dichos datos. Es importante señalar que este proyecto se encuentra conectado con la fuente oficial de datos que es generada de manera abierta por parte de la Organización Mundial de la Salud.

Al momento de navegar en CovidGraph, se obtiene acceso a datos de manera interactiva mediante una navegación entre nodos y aristas. En este sentido, los nodos representan datos en específico y las aristas la propiedad de la conexión entre los datos; es decir, el significado que tiene la unión entre datos de diferentes fuentes que están disponibles en el ambiente digital pero con un dominio en específico que está relacionado con la enfermedad del

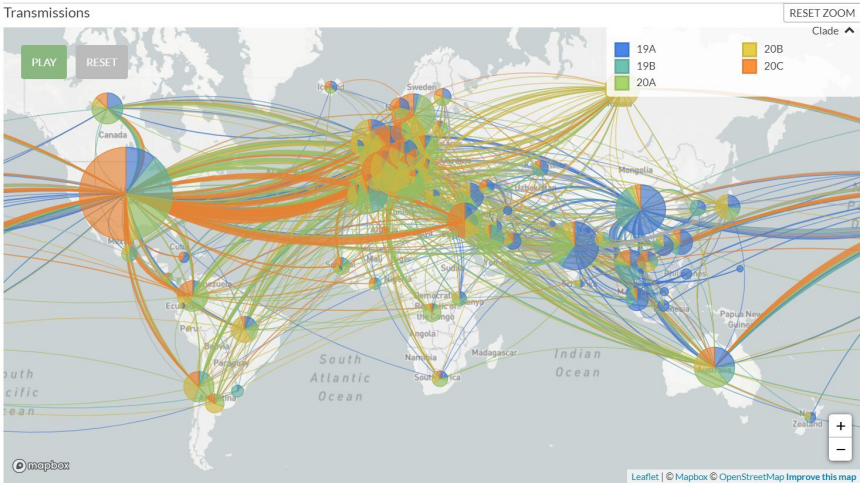
COVID-19. Este tipo de recuperación propicia el descubrimiento de patrones ocultos entre los mismos datos, pues mediante la interacción con el grafo el usuario puede descubrir nuevos hallazgos mediante un proceso intuitivo de búsqueda integradora y acorde a la demanda informativa que motiva a dicho usuario.

Por otra parte, el proyecto Linked COVID-19 Data del Robert Koch-Institut (http://covid19data.link/) presenta los resultados obtenidos de la implementación de una ontología (https://zenodo.org/record/3765375#.XtaE9jpKjb0) relacionada con la enfermedad del COVID-19. Además, ofrece una serie de visualizaciones que permiten obtener una consulta detallada del comportamiento de la enfermedad. Este proyecto ha sido desarrollado con datos abiertos liberados por la OMS y el centro de datos de la pandemia de la Johns Hopkins University. Estas fuentes han sido utilizadas para desarrollar cruces de datos y obtener estimaciones del impacto de la enfermedad sobre todo en los ámbitos sociales de las naciones.

Simultáneamente, la plataforma Genomic Epidemiology of Novel Coronavirus, generada por el proyecto Nextstrain, permite consultar en tiempo real el avance de la pandemia del coronavirus alrededor del mundo. Este proyecto está liberado bajo código abierto con el propósito de aprovechar el potencial de los datos y su aplicación en procesos científicos y de salud pública con miras a proporcionar una mejor comprensión de la pandemia y generar mejores respuestas al brote de la enfermedad. Para ello, han dado apertura a un grafo de conocimiento de índole geográfica con secuencia inmediata que permite visualizar el comportamiento y progreso de la pandemia (véase figura 5).

El grafo de Nextstrain está acompañado de una filogenia que muestra las relaciones evolutivas del virus SARS-CoV-2 y de la pandemia provocada por el coronavirus. Dicha filogenia muestra datos cronológicos referentes a la aparición inicial del coronavirus en Wuhan, China, en los meses de noviembre-diciembre de 2019, seguida de una transmisión sostenida de persona a persona que conduce a infecciones muestreadas y que permiten obtener una vista de la progresión de la pandemia.

Figura 5. Genomic Epidemiology of Novel Coronavirus

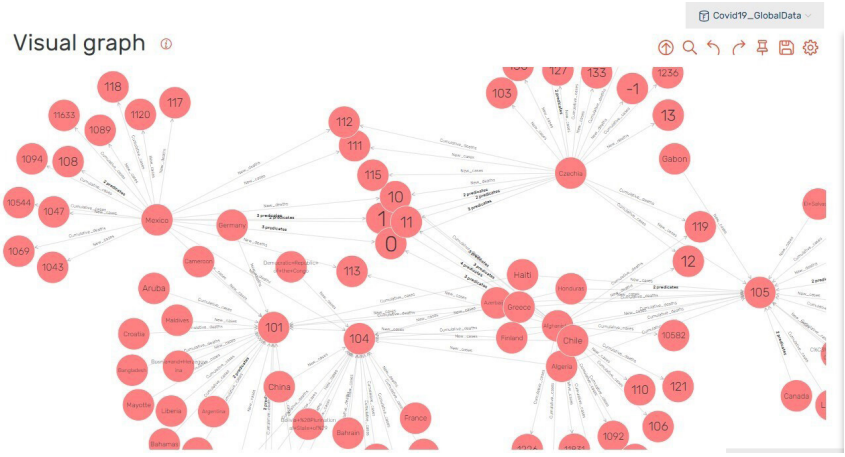


Fuente: https://nextstrain.org/ncov/global.

Se debe tener en cuenta que aunque las relaciones genéticas entre los virus muestreados son bastante claras, existe una considerable incertidumbre en torno a las estimaciones de fechas de transmisión específicas y en la reconstrucción de la propagación geográfica. Por lo tanto, hay que considerar que los patrones de transmisión inferidos específicos son solo una hipótesis y que los datos pueden aprovecharse para propósitos de reutilización y comprobación, pues el dinamismo de los datos representa los atributos cambiantes de la propagación de la pandemia.

Así pues, los datos abiertos de la pandemia remiten a fuentes de variada naturaleza, las cuales utilizan diferentes niveles de tecnología computacional y una amplia gama de formatos para estructurar y representar a sus datos. No existe un consenso en el uso de normatividad y estándares de codificación de los datos; sin embargo, el potencial de su visualización a gran escala permite estimar su aprovechamiento como fuente abierta mediante su descarga en formatos convencionales y ampliamente accesibles como CSV, XML y JSON.

Figura 6. COVID-19 data global



Fuente: elaboración propia con datos de la OMS, disponibles en https://covid19.who.int/.

Esta situación se ejemplifica en la figura 6, en donde puede apreciarse el desarrollo de un grafo de conocimiento utilizando los datos abiertos disponibles en la plataforma oficial de la Organización Mundial de la Salud. La fecha de corte de estos datos al momento de realizar esta representación fue el 2 de junio de 2020, cuando se habían reportado 6,194,533 casos confirmados de COVID-19, incluyendo 376,320 muertes reportadas a la OMS.

La reutilización de datos abiertos permite obtener representaciones que den la posibilidad para llevar a cabo el análisis de variables relacionadas con la pandemia, realizar comparaciones entre las cifras de diversos países y desarrollar estimaciones basadas en evidencias numéricas, todo ello en un solo dominio de datos. Dado el gran volumen y la alta complejidad de los datos de la pandemia, su visualización se emplea como un método para representar patrones y tendencias subyacentes, especialmente para audiencias que pueden carecer de experiencia para interactuar directamente con conjuntos de datos a gran escala o bien, para comprender las características del fenómeno.

Bajo esta premisa, los datos a menudo se pueden analizar de manera muy efectiva utilizando técnicas de visualización. Se estima que los grafos de conocimiento pueden aportar estrategias efectivas para el análisis y el acceso a enormes cantidades de datos, como es el caso de los generados por la pandemia. La investigación sobre el virus y las enfermedades generadas como parte de la pandemia pueden analizarse de manera integradora mediante este tipo de representaciones.

Sin embargo, será necesario continuar con la tendencia de la disponibilidad y apertura de los datos, pues como se ha mencionado anteriormente en este trabajo, los datos son la materia prima de la investigación y en tiempos de pandemia será relevante contar con datos significativos que ayuden a contrarrestar sus efectos devastadores.

Consideraciones finales

La investigación del SARS-CoV2 mediante el uso de datos abiertos y grafos de conocimiento permite obtener una visión de modelos para recuperar información. Estos modelos se caracterizan por el uso de métodos visuales e interactivos que son construidos mediante el uso de tecnología computacional, lo cual ha contribuido a la consulta y reutilización de los datos con el propósito de fomentar el descubrimiento de nuevos hallazgos en la investigación respecto a la pandemia, pues además de visualizar a los datos, es posible descargarlos para su latente reutilización.

En este sentido, la apertura de los datos y su respectiva visualización mediante grafos de conocimiento permite obtener un método de descubrimiento integrador, lo cual hace posible la consulta de diversos tipos de datos en un escenario de interacción común. Si bien el uso de grafo no es un tema novedoso para los campos de la ciencia de la salud y las ciencias de la computación, en tiempos actuales su uso es de relevancia para identificar comportamientos complejos entre los datos y la información que se han generado como parte del fenómeno de la pandemia.

Además, mientras no exista una vacuna o antiviral que cure los efectos de la COVID-19 aunado a la plena identificación del comportamiento y características infecciosas del virus SARS-CoV2, la humanidad seguirá conviviendo con la pandemia durante un largo periodo de tiempo, pues la nueva normalidad no será una constante sin la erradicación significativa de los patógenos que han cobrado la vida de miles de personas alrededor del mundo.

Por ello, las naciones deben comprender que la inversión en ciencia, tecnología y el desarrollo de políticas de datos abiertos pueden ser factores detonantes para el descubrimiento intensivo de nuevos hallazgos que permitan generar un antídoto. Además, la industria farmacéutica deberá socializar y humanizar los hallazgos para que el diseño de fármacos pueda devolver la salud a los enfermos de la pandemia.

REFERENCIAS

Charalabidis, Yanis *et al.*, "Open Data Interoperability", en *The World of Open Data: Public Administration and Information Technology*, vol 28. Suiza: Springer, 2018.

———. *The World of Open Data Concepts, Methods, Tools and Experiences*. Suiza: Springer, 2018.

Faniel, Ixchel y Trond Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues Data", *Computer Supported Cooperative Work* 19, (2010), 355-375, https://doi-org.pbidi.unam.mx:2443/10.1007/s10606-010-9117-8 (Consultado el 20 de mayo de 2020).

Giordano, Paolo. *En tiempos de contagio*. España: Salamandra, 2020.

Kale, Nitin y Nancy Jones, *Practical Analytics*. Epistemy Press, 2020.

Kejriwal Mayank. "What Is a Knowledge Graph?", en *Domain-Specific Knowledge Graph Construction*, editado por Mayank Kejriwal, 1-7. Suiza: Springer International Publishing, 2019.

Mira, J. "Pandemia COVID-19: y ahora ¿qué?", *Journal of Healthcare Quality Research* 785 (2020), https://doi.org/10.1016/j.jhqr.2020.04.001 (Consultado el 27 de mayo de 2020).

Mundie, Craig. "El camino por recorrer". En *El cuarto paradigma: descubrimiento científico intensivo de datos*, editado por Tony Hey, Stewart Tansley y Kristin Tolle, 241-244. México: Universidad Autónoma Metropolitana, 2009.

Musa Aliyu y Adegboyega Ojo, "Towards Building a Knowledge Graph with Open Data – A Roadmap", en International Conference on e-Infrastructure and e-Services for Developing Countries, 157-162. Lagos: EAI, 2017.

Nicole, Lurie, Melanie Saville, Richard Hatchett, y Jane Halton. "Developing COVID-19 Vaccines at Pandemic Speed". *New England Journal of Medicine* 382, no. 21 (2020), https://doi.org/10.1056/NEJMp2005630 (Consultado el 22 de mayo de 2020).

Organización Mundial de la Salud, "Preguntas y respuestas sobre la enfermedad por coronavirus (COVID-19)", https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses (Consultado el 20-05-2020).

Tenopir, Carol. *et al.*, "Data Sharing by Scientists: Practices and Perceptions", *PLoS ONE*, no. 6 (2011), https://doi.org/10.1371/journal.pone.0021101 (Consultado el 28 de mayo de 2020).

Wallis Jilian, Elizabeth Rolando y Christine Borgman. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology", *Plos One (2013),* https://doi.org/10.1371/journal.pone.0067332 (Consultado el 21 de mayo de 2020).

Zenodo, "Coronavirus Disease Research Community COVID-19", https://zenodo.org/communities/covid-19/?page=1&size=20 (Consultado el 1 de junio de 2020).