Virus de ácido ribonucleico (ARN) y coronavirus en *Google Dataset Search*: alcance y correlación epidemiológica

Ribonucleic acid (RNA) virus and coronavirus in *Google Dataset Search*: their scope and epidemiological correlation

Manuel Blázquez-Ochando; **Juan-José Prieto-Gutiérrez**

Cómo citar este artículo:

Blázquez-Ochando, Manuel; **Prieto-Gutiérrez, Juan-José** (2020). "Virus de ácido ribonucleico (ARN) y coronavirus en *Google Dataset Search*: alcance y correlación epidemiológica". *Profesional de la información*, v. 29, n. 6, e290628.

https://doi.org/10.3145/epi.2020.nov.28

Artículo recibido el 24-05-2020
Aceptación definitiva: 21-10-2020

Manuel Blázquez-Ochando ✉
https://orcid.org/0000-0002-4108-7531

Universidad Complutense de Madrid
Facultad de Ciencias de la Documentación
Santísima Trinidad, 37
28010 Madrid, España
manublaz@ucm.es

Juan-José Prieto Gutiérrez
https://orcid.org/0000-0002-1730-8621

Universidad Complutense de Madrid
Facultad de Ciencias de la Documentación
Santísima Trinidad, 37
28010 Madrid, España
jujpriet@ucm.es

Resumen

Se presenta un análisis sobre la publicación de conjuntos de datos recogidos en el buscador *Google Dataset Search*, especializados en familias de virus de ARN, cuya terminología fue obtenida en el tesauro del *National Cancer Institute* (*NCI*), elaborado por el *Department of Health and Human Services* de los Estados Unidos. Se busca evaluar el alcance y capacidad de reutilización de los datos disponibles, determinando el número de datasets, su libre acceso, proporción en formatos de descarga reutilizables, principales proveedores, cronología de publicación y verificación de su procedencia científica. Por otra parte, definir posibles vínculos entre la publicación de datasets y las principales pandemias ocurridas en los últimos 10 años. Entre los resultados obtenidos se destaca que sólo el 52% de los datasets tienen correspondencia con investigaciones científicas y, en menor medida, un 15% son reaprovechables. También se observa una evolución al alza en la publicación de datasets, especialmente vinculada a la afectación de las principales epidemias. Esto es confirmado de manera evidente con los virus del Ébola, Zika, SARS-CoV, H1N1, H1N5 y, particularmente con el coronavirus SARS-CoV-2. Finalmente, se observa que el buscador aún no ha implementado métodos adecuados para el filtrado y supervisión de los datasets. Estos resultados muestran algunas de las dificultades que aún presenta la ciencia abierta en el campo de los datasets.

Palabras clave

Datos; Datasets; Conjuntos de datos; Virus; Virus de ARN; Coronavirus; SARS-CoV-2; Covid-19; Pandemias; Reutilización de datos; *Google*; *Google Dataset Search*; Proveedores de datos; Buscadores; Recuperación de información; Ciencia abierta.

Abstract

This paper presents an analysis of the publication of datasets collected via *Google Dataset Search*, specialized in families of RNA viruses, whose terminology was obtained from the *National Cancer Institute* (*NCI*) thesaurus developed by the US *Department of Health and Human Services*. The objective is to determine the scope and reuse capacity of the available data, determine the number of datasets and their free access, the proportion in reusable download formats, the main providers, their publication chronology, and to verify their scientific provenance. On the other hand, we also define possible relationships between the publication of datasets and the main pandemics that have occurred during the last 10 years. The results obtained highlight that only 52% of the datasets are related to scientific research, while an even

smaller fraction (15%) are reusable. There is also an upward trend in the publication of datasets, especially related to the impact of the main epidemics, as clearly confirmed for the Ebola virus, Zika, SARS-CoV, H1N1, H1N5, and especially the SARS-CoV-2 coronavirus. Finally, it is observed that the search engine has not yet implemented adequate methods for filtering and monitoring the datasets. These results reveal some of the difficulties facing open science in the dataset field.

Keywords

Data; Datasets; Viruses; RNA viruses; Coronavirus; SARS-CoV-2; Covid-19; Pandemics; Data reuse; *Google*; *Google Dataset Search*; Data providers; Search engines; Information retrieval; Open science.

1. Introducción

La consulta de conjuntos de datos (datasets)[1] y de artículos ubicados en repositorios se ha convertido en práctica habitual y papel central en la investigación (**Marcial**; **Hemminger**, 2010) para la toma de decisiones bien fundamentadas (**Hernández-Pérez**, 2016). Por ejemplo, el tamaño medio anual del conjunto de datos en los artículos de *Miccai* (*Medical Image Computing and Computer-Assisted Intervention*) ha crecido aproximadamente de 3 a 10 veces entre 2011 y 2018 (**Landau**; **Kiryati**, 2019), lo que viene a confirmar esta orientación y pone de manifiesto el cambio de paradigma hacia una ciencia abierta.

Cada día hay mayor conciencia de la necesidad de compartir datos y materiales derivados de las investigaciones científicas para reproducir análisis, compararlos y plantear nuevas preguntas (**Nosek** *et al*., 2015), aunque es cierto que hay ciertas inquietudes sobre la confidencialidad, gobernanza y posible uso indebido de datos institucionales y comerciales (**Howe** *et al.*, 2018). Sin embargo se considera que los inconvenientes son inferiores a los beneficios, según afirma una encuesta a 800 investigadores, en donde se concluía que menos del 8% consideró las posibles consecuencias negativas del intercambio de datos (**Mello**; **Lieou**; **Goodman**, 2018).

Los últimos trabajos de datos abiertos (**Corrales-Garay**; **Ortiz-de-Urbina-Criado**; **Mora-Valentín**, 2019), muestran un aumento de su uso debido a:

- cambio de conducta en las investigaciones científicas, que ahora también se asientan en el análisis masivo de datos o *big data* (**Saheb**; **Izadi**, 2019);
- control de las leyes de protección de datos (**Polonetsky**; **Tene**; **Finch**, 2016);
- aumento de transparencia (**Weston** *et al*., 2019);
- exigencias de los organismos financiadores, que obligan a sustentar las conclusiones científicas en datos probados y reconocibles en datasets compartidos.

Esto se puede comprobar en el *Plan S* de la *Comisión Europea* (*Science Europe*, 2019) en el que se insta, desde el 1 de enero de 2021, a la publicación en revistas de acceso abierto dorado o en repositorios y "plataformas afines" que publican pdfs editoriales, siendo además una gran oportunidad para que las revistas lleven a cabo una transformación digital plena (**López-Borrull** *et al*., 2020).

Recientemente los preprints científicos en acceso abierto se están convirtiendo en una fuente de información fundamental para enfrentar cuestiones trascendentales, como la crisis sanitaria producida por el SARS-CoV-2 (**Johansson**; **Saderi**, 2020), en la que investigadores de todo el mundo están uniendo sus esfuerzos, conocimientos y bases de datos para:

- identificación de pacientes infectados mediante la sintomatología de la fiebre y su patrón de incidencia (**Haleem** *et al*., 2020);
- predicción espacio-temporal de velocidad y magnitud de la transmisión del virus (**Zhou** *et al*., 2020);
- simulación del plegamiento de proteínas para terapias dirigidas (**Chen** *et al*., 2020);
- predicción del progreso de la enfermedad Covid-19 a través de imágenes radiológicas (**Chen**; **Lerman**; **Ferrara**, 2020);
- obtención de tratamientos y vacunas efectivas (**Le-Guillou**, 2020).

El objeto de este trabajo es el análisis de los resultados obtenidos en el buscador de datasets de *Google* en referencia a familias de virus, para lo que se plantean las siguientes preguntas de investigación:

¿*Google Dataset Search* es un buscador de datos abiertos, adecuado para la investigación?

¿Cuál es la proporción de datasets abiertos y reutilizables para las consultas de virus preestablecidas?

¿Qué virus ARN presentan un mayor número de datasets de investigación?

¿Cuál es la evolución de los datasets en función del virus?

¿Existe correlación entre la cronología de las epidemias y la publicación de datasets y artículos científicos?

¿Qué dificultades documentales se encuentran en los datasets científicos abiertos?

¿Cómo pueden reutilizarse los datasets obtenidos?

Las cuestiones de investigación se orientan en una doble vertiente de Información-Documentación y correlativa de virus. Esto es así porque la cantidad de datasets y su evolución está relacionada entre otras cuestiones con las afecciones víricas, tal como se explica en la investigación.

El buscador *Google Dataset Search* es una utilidad de consulta especializada en conjuntos de datos, que recopila información procedente de repositorios científicos, comerciales y gubernamentales de muy diversa índole, tal como se explicará más adelante. Sus características no son comparables a otros buscadores y agregadores, ni tan siquiera con los proveedores de datos a los que da cobertura en un corpus único.

> Un dataset o conjunto de datos es una colección de datos estructurada y delimitada en sus valores, de forma que puede ser reaprovechada en bases de datos, hojas de cálculo, programas de análisis estadístico y *big data*

Google recopila para su buscador datasets que cumplen el estándar de metadatos *Schema.org* (**Brickley**; **Burgess**; **Noy**, 2019), más concretamente el referido a datasets, según se explica en su enfoque sobre el descubrimiento de conjuntos de datos (*Google*, 2020). Esto ha permitido crear un instrumento de recopilación masiva que sin duda facilita la investigación científica. Sin embargo, también plantea dudas sobre su idoneidad, como se explicará más adelante. Este buscador entró en servicio el 5 de septiembre de 2018, pero no estuvo abierto definitivamente hasta enero de 2020, momento en el que ha coincidido con la crisis del coronavirus, por lo que puede demostrar su versatilidad, ventajas y carencias, en una situación de urgencia y necesidad. Adicionalmente, en esta investigación se trata de observar las limitaciones del buscador, así como la corrección de la información presentada, sus posibilidades de reutilización y pertinencia, efectuando búsquedas especializadas en virología.

2. Metodología

La metodología empleada en este trabajo tiene por objeto la preparación de una muestra de consultas a partir de un vocabulario controlado, que permita interrogar al buscador *Google Dataset Search*. Este método es empleado habitualmente en los trabajos de evaluación de buscadores (**Lewandowski**, 2015), a fin de observar la relevancia y pertinencia de los resultados, al margen de otros métodos de evaluación (**Hawking** *et al.*, 2001). Para ello, se requiere la selección de una muestra de vocabulario, procedente de un lenguaje documental, esto es, un vocabulario controlado y normalizado, a fin de asegurar la repetibilidad y reproducibilidad de las pruebas sobre el buscador, según **Broder** (2002). Esto también es representativo de la metodología TREC de consulta y evaluación (**Hawking** *et al.*, 1999). Cada palabra o sintagma nominal elegido sirve para componer una consulta, para lo cual se diseña una fórmula de consulta cuya complejidad variará en función del propósito de la evaluación. En esta investigación, según se explicará más adelante, no se requieren ecuaciones complejas, bastando con la búsqueda de términos literales, en la variable de consulta por defecto del buscador. Esto permite elaborar un url de búsqueda que puede ser ejecutado de forma manual o automática, con objeto de recopilar los resultados que arroja el buscador. En esta investigación se observan:

- los valores cuantitativos del número total de resultados de la consulta;
- la proporción de los resultados según el tipo de acceso y formato;
- la cronología de publicación y registro de los datasets;
- el número de artículos científicos vinculados a los datasets recuperados;

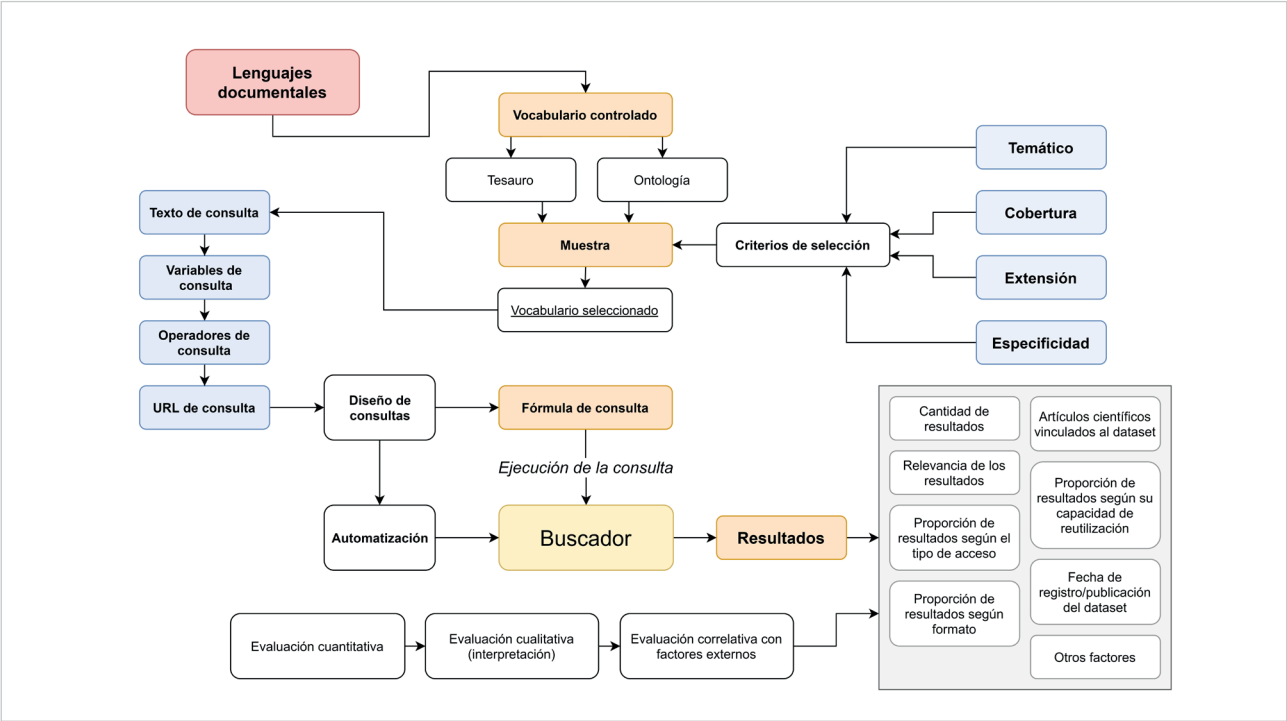


Figura 1. Modelo metodológico de evaluación de buscadores, empleado en la investigación

- la procedencia de los conjuntos de datos (estos son los principales proveedores de datos);
- la proporción de conjuntos de datos reutilizables para tareas de investigación y documentación;
- la relevancia de los resultados conforme a cada consulta planteada (esto es, la adecuación de los resultados al término del vocabulario empleado) o lo que es lo mismo, el análisis del contenido que ofrece el buscador para cada consulta (**King** *et al.*, 2007).

A continuación, se profundiza en las fases del método: a) selección de terminología especializada; b) configuración y ejecución de consultas en *Google Dataset Search*; c) análisis de resultados.

- Para la selección de terminología especializada en familias de virus se ha elegido el tesauro del *National Cancer Institute* (*NCI*), que es utilizado extensivamente por los *National Institutes of Health* (*NIH*). A continuación se ha seleccionado la sección del tesauro que contiene las referencias terminológicas de virus (Código C14283), entre las que figuran virus de ADN, retrovirus, virus de ARN y otros grupos de virus. Para esta investigación, se han elegido los virus de ARN por encontrarse entre ellos los de la familia Coronaviridae, que han ocasionado la alerta sanitaria internacional en 2020. De esta forma también se podrá comparar cuantitativamente el número de datasets recuperados del SARS-CoV-2 con respecto al resto de virus de la misma clasificación. En suma, se han obtenido 22 familias de virus de tipo ARN y 70 términos, que pueden comprobarse en la tabla 1.

Tabla 1. Términos obtenidos del tesauro del *NCI* que han sido utilizados para efectuar las consultas en *Google Dataset Search*

Familias de virus (Q1)	Virus (Q2)
Reoviridae	Colorado tick fever virus, Orbivirusm, Rotavirus
Arenavirus	Lymphocytic choriomeningitis virus, Tacaribe virus
Bunyaviridae	Hantavirus, Nairovirus, Orthobunyavirus, Phlebovirus
Filoviridae	Ebola virus, Marburgvirus
Influenza	Avian influenza, H5N2 avian influenza, Influenza H1N1, Influenza H5N1
Avulavirus	Newcastle disease virus
Henipavirus	Hendra virus, Nipah virus
Morbillivirus	Measles morbillivirus
Paramyxovirus	Human parainfluenza
Respirovirus	Human parainfluenza virus 1, Human parainfluenza virus 3
Rubulavirus	Human parainfluenza virus 2, Human parainfluenza virus 4, Mumps virus
Pneumovirinae	Human respiratory syncytial virus, Metapneumovirus, Pneumovirus
Rhabdoviridae	Rabies virus
Astroviridae	Astroviridae
Hepatitis E	Hepatitis E
Norovirus	Norovirus genogroup, Norwalk virus
Sapovirus	Sapporo Virus
Coronavirus	Porcine epidemic diarrhea virus, SARS Coronavirus, SARS Coronavirus 2, Covid-19, Covid-20, Covid-21, Covid-22, Covid-23, Deltacoronavirus, Gammacoronavirus
Flavivirus	Dengue virus, Powassan virus, Tick-Borne encephalitis virus, West Nile virus, Yellow fever virus, Zika virus, Hepatitis C virus, Pegivirus
Picornavirus	Aphthovirus, Coxsackie A virus, Coxsackie B virus, Echovirus, Enterovirus D68, Enterovirus D69, Enterovirus D70, Enterovirus D71, Enterovirus D72, Poliovirus, Hepatitis A virus, Rhinovirus
Togaviridae	Togaviridae
Alphavirus	Barmah forest virus, Chikungunya virus, Ross river virus, Rubella virus, Rubivirus

- Configuración y ejecución de consultas en *Google Dataset Search*. Se han considerado dos tipos de consultas. En primer lugar, búsquedas de las familias de virus, a las que se denomina en clave consultas Q1 y, en segundo lugar, búsquedas de las especies de virus, que se denominan consultas Q2. Se obtienen los principales proveedores de datos, formatos de los datasets, su reutilización, tipo de acceso y la cronología de su publicación para cada especie de virus. Es relevante precisar que un resultado dado por el buscador puede suponer uno o más datasets, ya que a menudo pueden encontrarse en diversos formatos de exportación. En este trabajo se ha considerado específicamente el número de datasets, para obtener una mayor precisión en las cifras aportadas. El modo de consulta se basa en el uso directo de los descriptores del tesauro, a los que se añaden comillas de búsqueda literal, por ejemplo “Porcine Epidemic Diarrhea Virus”, a fin de obtener los conjuntos de datos más pertinentes a cada virus en cuestión.

- Análisis de resultados. El proceso de análisis tiene por objeto comparar los datos de las consultas Q1 y Q2. Estos datos se comparan con una muestra de Q2, configurada con los datasets de los 5 primeros resultados de cada especie de virus, para su análisis prospectivo. Se observan posibles diferencias cuantitativas y cualitativas, su frecuencia de publicación, correlación con las principales epidemias y pandemias, evolución en los últimos 20 años y su comparación con la publicación de artículos científicos. Adicionalmente se apuntan las limitaciones y problemáticas observadas en *Google Dataset Search*.

3. Resultados

Las consultas de categoría Q1, relativas a familias de virus, arrojan un total de 1.375 datasets, de los cuáles casi un 87% son de acceso abierto. Al filtrar los resultados por formatos concretos, aptos para la reutilización de la información, esto es, formatos CSV (valores separados por comas) y XLS (hoja de cálculo *Excel* o similar), sólo el 58,59% se pueden aprovechar en bases de datos y programas de análisis estadístico, puesto que el 41,41% restante se encuentran en formatos de imagen o documentos de tipo pdf u ofimáticos.

Los resultados de la categoría Q2 son de proporciones similares a las ya advertidas en Q1. El 82% de los resultados corresponden a datasets de libre acceso y el 18% restante a conjuntos de datos bajo suscripción. Por otra parte, el 64% se encuentra en formatos aprovechables y el 36% en formato de imagen u ofimáticos derivados. Estos datos preliminares confirman el auge de la ciencia abierta en este sector.

Sin embargo, al examinar detenidamente la muestra de Q2 se observan resultados bien diferentes, que plantean serias dudas con respecto a la información que a priori proporciona *Google Dataset Search*. De los 3.799 conjuntos de datos obtenidos, se analizan en profundidad los de los 5 primeros resultados más relevantes para cada especie de virus, obteniendo una muestra de 331 datasets. A partir de su escrutinio se llegó a la conclusión de que sólo 67 son reutilizables, al estar disponibles en formatos estándar CSV, XLS, SQL y XML. Esto supone que sólo un 21% del total es aprovechable, lo que significa una discordancia con la información ofrecida por el buscador en las consultas Q1 y Q2.

Tabla 2. Resultados obtenidos para las consultas Q1, Q2 y muestra de Q2

	Q1		**Q2**		**Muestra Q2**	
	Número de datasets	%	Número de datasets	%	Número de datasets	%
Número total de datasets recuperados	**1.375**	**100%**	**3.799**	**100%**	**331**	**100%**
Distribución según el acceso						
Acceso abierto	1.197	86,80%	2.890	76,07%	317	95,77%
Suscripción o pago	182	13,20%	641	23,93%	14	4,22%
Distribución según la reutilización						
Reutilizables (.csv, .xls, .sql, .xml)	808	58,59%	2.423	63,78%	71	21,45%
No reutilizables (.pdf, .doc, .ppt, .png, .tif, .jpg)	571	41,41%	1.376	36,22%	260	78,55%

Fuente: *Google Dataset Search*. Fecha de recopilación: 8 de marzo de 2020

> La Covid-19 asociada al SARS-CoV-2 presenta 94 conjuntos de datos en sólo 3 meses, lo que supera cualquier previsión

Profundizando en los datos obtenidos, en la tabla 3 se observa que el número de conjuntos de datos reutilizables se reduce al 15% si se añade el criterio de procedencia científica demostrada. Esto es, que al menos se compruebe el vínculo entre el dataset y un artículo científico. También se ha descubierto que el 37% de los datasets no reutilizables tenían procedencia científica, correspondiendo en su mayoría a imágenes, ilustraciones y *papers* científicos, lo que significa que en realidad no son conjuntos de datos. Por otra parte, también se encuentra un margen del 6% de los datasets reutilizables que no tenían un origen científico acreditado, por carecer de investigación adjunta. Además, el 66% de los conjuntos de datos publicados se concentran en los últimos 5 años, lo que indica que la frecuencia de actualización de los datos puede ser baja. A estas cifras cabe añadir el factor de agregación y heterogeneidad de los datos, ya que en muchos datasets similares, el objeto del contenido, la estructura y variables no coincidían, lo que significa que no son continuados en la mayoría de los casos. Por consiguiente, el factor de agregación y continuidad de los mismos resulta muy bajo.

En la figura 2 se presenta la frecuencia de publicación absoluta de los conjuntos de datos de cada virus. Su número no presenta una progresión lineal hasta el año 2010, momento en el que se produce un crecimiento constante y sostenido. En 2016 se recupera el mayor número de conjuntos de datos relativos al ensayo de virus, destacando:

- el *Lymphocytic Choriomeningitis* o virus de la Coriomeningitis, producida por roedores (29 resultados);
- el virus Nipah endémico de la zona de Malasia e India (31 resultados);
- el Rhinovirus o virus del resfriado común (25 resultados);
- el SARS Coronavirus o SARS-CoV (25 resultados).

Tabla 3. Distribución cronológica de los datasets analizados en la muestra de Q2

Fecha	Datasets analizados	Científicos	No científicos	Reutilizables (.csv, .xls, .sql)	No reutilizables (No son data-sets)	Reutilizables de origen científico	No reutilizables de origen científico	Reutilizables sin origen científico acreditado
< 2010	87	4	89	1	86	1	3	0
2010	1	0	1	0	1	0	0	0
2011	1	0	0	1	0	0	0	0
2012	10	2	5	0	10	0	2	0
2013	12	4	8	0	12	0	4	0
2014	19	3	16	2	17	1	2	1
2015	34	29	5	13	21	11	18	2
2016	55	54	1	11	44	11	43	0
2017	16	12	3	5	11	4	8	1
2018	28	22	6	9	19	6	16	3
2019	40	30	9	10	30	7	23	3
2020	28	12	16	19	9	9	3	10
Suma	331	172	159	71	260	50	122	20
%	100%	52%	48%	21%	79%	15%	37%	6%

A pesar del descenso en el número de datasets publicados en 2017, con un índice similar a 2014, se observa una vuelta al crecimiento hasta la fecha, que confirma una disposición al alza. De hecho, en el año actual 2020 se registra el mayor número asociados a un virus. La Covid-19 asociada al SARS-CoV-2 presenta 94 conjuntos de datos en sólo 3 meses, lo que supera cualquier previsión.

En este estadio de la investigación cabe preguntarse si hay una relación de causa-efecto entre los principales episodios de epidemias y pandemias que han trascendido en la bibliografía científica y los medios de comunicación, en referencia a los conjuntos de datos. Para responder a esta pregunta se ha elaborado la figura 3, en la que se superponen a la línea cronológica las epidemias y pandemias más conocidas con respecto a las fechas de publicación de conjuntos de datos

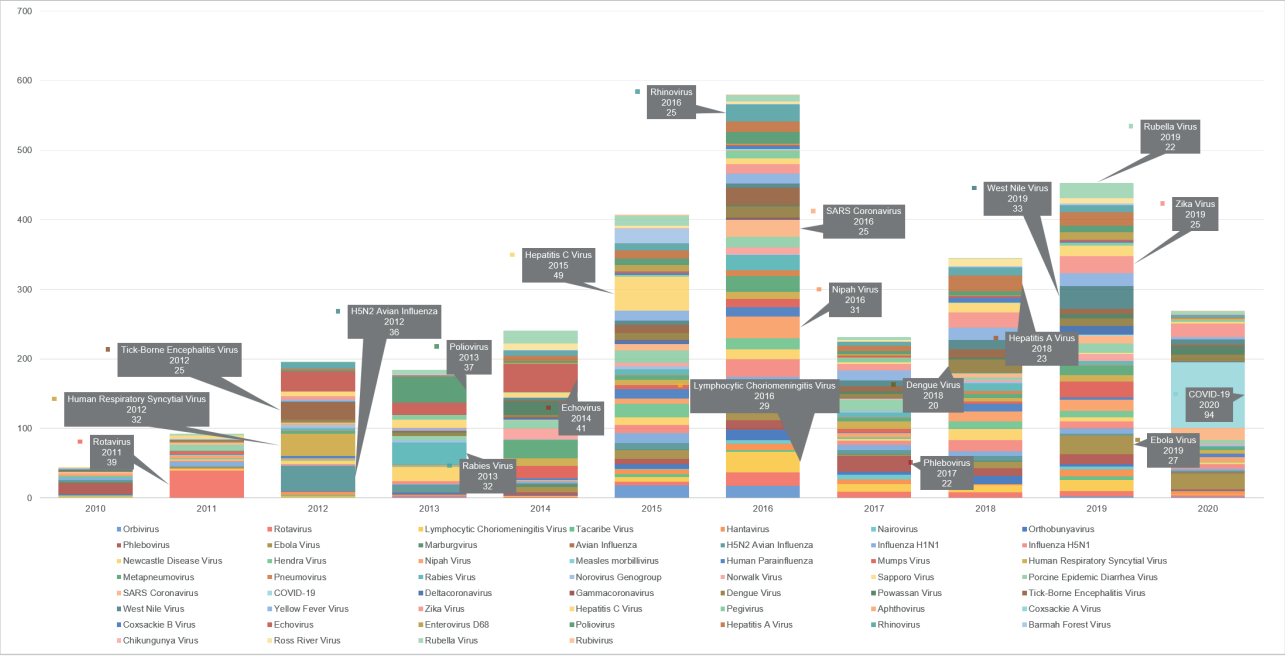
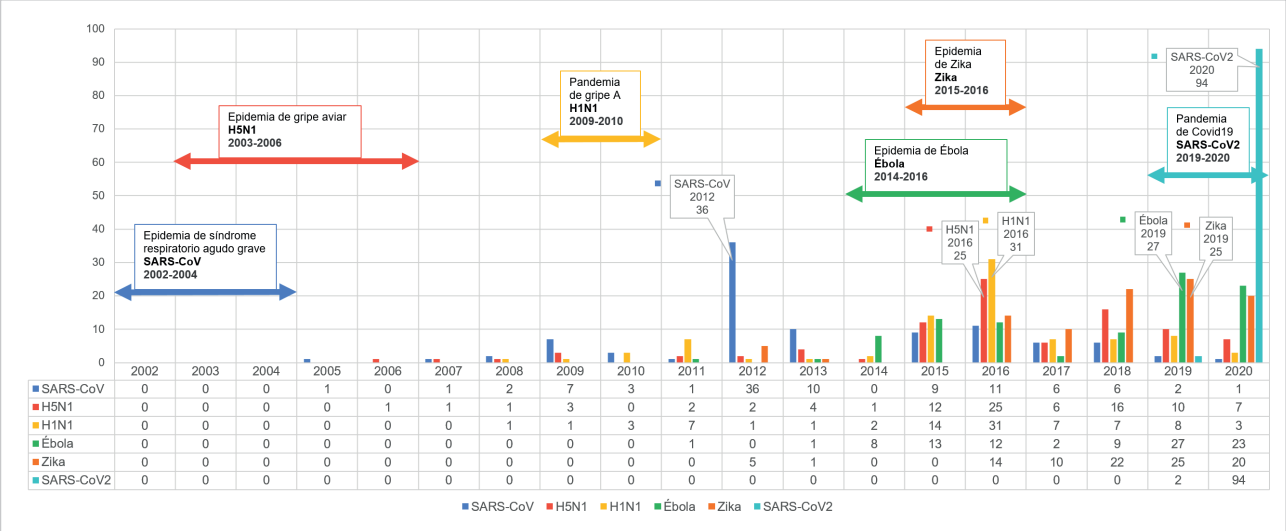


Figura 2. Publicación de datasets de virus ARN en los últimos 10 años. Fuente: *Google Dataset Search*. Fecha de recopilación: 8 mayo 2020. Datos disponibles en:
https://github.com/manublaz/datasets/blob/master/cronologiaPublicacionDatasetsVirusARN.xlsx



	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
SARS-CoV	0	0	0	1	0	1	2	7	3	1	36	10	0	9	11	6	6	2	1
H5N1	0	0	0	0	1	1	1	3	0	2	2	4	1	12	25	6	16	10	7
H1N1	0	0	0	0	0	0	1	1	3	7	1	1	2	14	31	7	7	8	3
Ébola	0	0	0	0	0	0	0	0	0	1	0	1	8	13	12	2	9	27	23
Zika	0	0	0	0	0	0	0	0	0	0	5	1	0	0	14	10	22	25	20
SARS-CoV2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	94

Figura 3. Cronología de epidemias y publicación de datasets.
Fuente: *Google Dataset Search*. Fecha de recopilación: 8 mayo 2020. Datos disponibles en:
https://github.com/manublaz/datasets/blob/master/cronologiaPublicacionDatasetsPapersEpidemias.xlsx

registradas en *Google Dataset Search*. Los resultados indican que siempre se produce un retraso en la publicación de datasets con respecto a las fechas de los brotes o epidemias, siendo el caso más notable el del SARS-CoV que tuvo lugar entre 2002 y 2004. Si bien, el primer conjunto de datos sobre SARS-CoV queda registrado al año siguiente al término de la epidemia, los valores cuantitativos son casi intrascendentes hasta 2012, cuando se publica la mayoría de los datos, un total de 36 datasets.

Aunque como se ha indicado anteriormente, sólo el 52% de los datos son científicos y sólo el 15% reutilizables, se puede afirmar que la gravedad de la crisis de coronavirus ha provocado una rápida respuesta de la comunidad científica que ha tenido un reflejo elocuente en las estadísticas, aunque aún no pueda confirmarse el grado de coordinación, agregación y agrupación de los datos según los diversos sujetos productores.

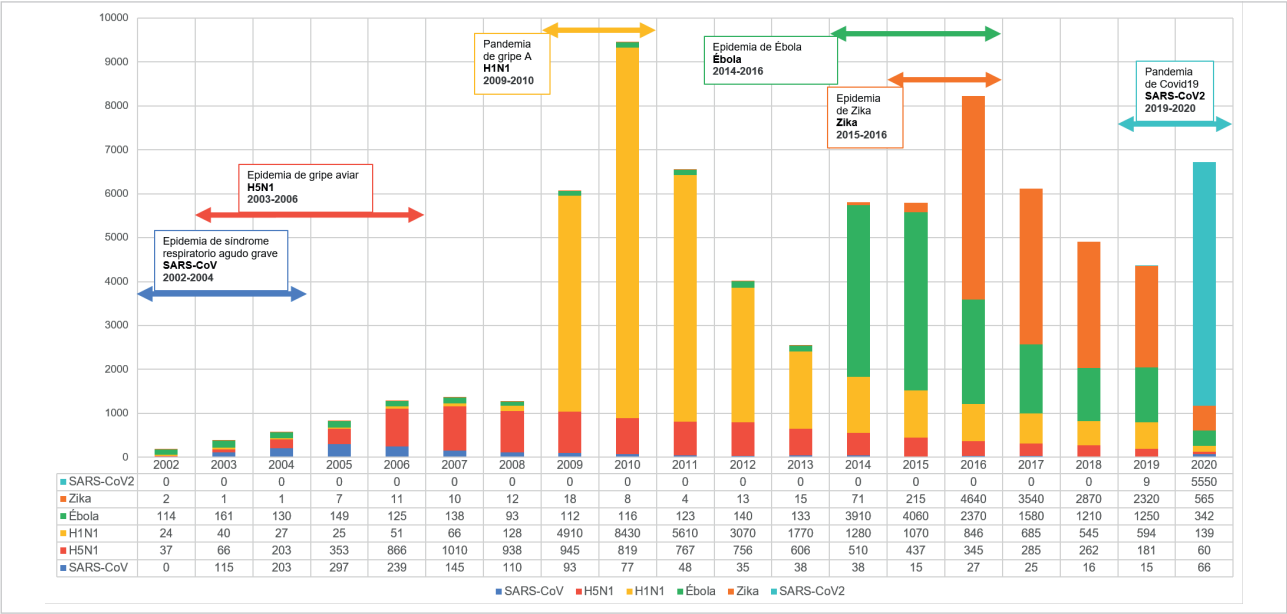
Este patrón se repite en la mayoría de los casos, por ejemplo, en la epidemia de gripe aviar provocada por el H5N1 entre 2003 y 2006 y la gripe A o del H1N1 entre 2009 y 2010. En ambos casos, las cifras de datasets publicados se mantienen por debajo de 10 hasta 2016, cuando se produce el mayor incremento de la serie histórica. Otro caso destacable es la epidemia de Ébola entre 2014 y 2016, de la que sí se observan datasets publicados antes y durante el brote. Esto puede explicarse fácilmente, ya que el virus del Ébola es conocido desde 1976 (**Emond** *et al.*, 1977), fecha desde la que se han registrado al menos siete brotes. Este factor, unido a la alta mortalidad y morbilidad, han podido influir en el interés de la comunidad científica. De hecho, se produce un incremento en el número de trabajos y la cantidad de conjuntos de datos publicados, que alcanza su máximo en 2019.

Otro caso similar es el del virus Zika. Si bien se hallan datasets desde 2012, la epidemia de Zika no tuvo lugar hasta 2015 y 2016, momento en el que el brote epidémico alcanzó América del sur, central y parte del Caribe. Al igual que el Ébola, el Zika era conocido con antelación, desde 1947 (**Dick**; **Kitchen**; **Haddow**, 1952), lo que explica la existencia de datasets previos a la última epidemia. Sin embargo, sí se observa un crecimiento de conjuntos de datos desde entonces, que tiene sus valores máximos entre 2019 y 2020. Si los resultados obtenidos para los datasets se los compara con los artículos científicos publicados, se observa que su frecuencia de publicación es mayor y se solapan a las fechas de incidencia de las pandemias (figura 4). Este fenómeno está claramente justificado por ser el medio de comunicación científica preferente y ayuda a poner en contexto, tanto en volumen como en periodicidad, la publicación de artículos y datasets, aunque su ratio comparativa sea extremadamente baja.

Nunca antes se había alcanzado este nivel de preocupación, reflejado en publicaciones científicas y datasets, lo que demuestra el grado de atención de la comunidad científica ante las dificultades esenciales que afectan a la sociedad en su conjunto. Otros aspectos para destacar son la aparición de datasets sobre SARS-CoV-2, justo al inicio de la crisis del coronavirus en diciembre de 2019, así como el salto cuantitativo producido en tan reducido espacio de tiempo.

Los resultados obtenidos sugieren varias reflexiones sobre los factores que afectan a la producción de los datasets especializados en virología, entre los cuales podrían destacarse los siguientes:

- Gravedad de la pandemia. Se deduce que a mayor afectación y peligrosidad de un virus, mayor será la cantidad de conjuntos de datos publicados, vinculados a investigaciones y a artículos científicos. Esto queda confirmado con el virus SARS-CoV-2, al demostrarse que la tasa de publicación de datasets por artículo es la más alta en comparación con el resto de virus analizados.
- Transparencia de las investigaciones. Un porcentaje notable de los datasets de investigación no son de libre acceso, lo que dificulta la comprobación independiente de los hechos científicos de los que se derivan las publicaciones científ-



	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
SARS-CoV2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	5550
Zika	2	1	1	7	11	10	12	18	8	4	13	15	71	215	4640	3540	2870	2320	565
Ébola	114	161	130	149	125	138	93	112	116	123	140	133	3910	4060	2370	1580	1210	1250	342
H1N1	24	40	27	25	51	66	128	4910	8430	5610	3070	1770	1280	1070	846	685	545	594	139
H5N1	37	66	203	353	866	1010	938	945	819	767	756	606	510	437	345	285	262	181	60
SARS-CoV	0	115	203	297	239	145	110	93	77	48	35	38	38	15	27	25	16	15	66

Figura 4. Cronología de epidemias y publicación de artículos científicos.
Fuente: *Google Scholar*. Fecha de recopilación: 8 mayo 2020.

ficas (**Irwin**, 2009). Los principios de reproducibilidad y repetibilidad quedan comprometidos por los intereses de las compañías que financian las investigaciones y minerías de datos (**Bekelman**; **MPhil**; **Gross**, 2003).

- Dificultades propias de la investigación y la integridad de los datos. Este factor podría explicar el lapso de tiempo que se produce entre una pandemia y la publicación de los datasets, tal como se viene describiendo. Ello puede deberse a la dificultad para ingeniar métodos y ensayos científicos adecuados, validados por la comunidad científica, las carencias en los procesos de automatización y recolección de los datos, o bien el prolongado proceso de verificación y curación de estos, propio de esta área de conocimiento (**Schneier**, 2012).

Por último, para precisar el alcance de la información analizada se revisan los proveedores que colaboran con *Google*, en donde se observan ciertas limitaciones. Una de las más relevantes es la falta de cifras oficiales sobre su cobertura. Por tanto, la información analizada sobre proveedores se corresponde con los 100 primeros resultados que recupera el buscador, en base a la muestra de familias de virus Q1, especies de virus Q2 y muestra de Q2.

A partir de aquí se puede afirmar que la cobertura de los conjuntos de datos y archivos recuperados sobre virus de ARN varía en función del proveedor. No todos los proveedores de datos suministran datasets aprovechables a *Google*, tal como se viene indicando. Incluso algunos proveedores de datos muestran archivos ofimáticos, que no tienen utilidad para conformar bases de conocimiento para *big data*. Sin embargo, desde el buscador no se hace distinción o valoración cualitativa a este respecto.

Según se muestra en la tabla 4, destacan dos proveedores de datos sobre el resto: *Figshare* y *ResearchGate*. En cuanto a la búsqueda general de familias de virus (Q1) ambos superan el 70% de cobertura y sobre las búsquedas específicas de cada familia (Q2) se acercan al 60%. El resto de proveedores analizados (*Statista*, *Omicsdi*, *Datamed*, *Search.datacite*, etc.) tienen una cobertura muy baja, por debajo del 4% individualmente, pero globalmente acumulan algo menos del 30% en Q1 y del 40% en Q2, aproximadamente.

Tabla 4. Principales proveedores de datos de los datasets analizados

Proveedores de datos	Q1		Q2		Muestra de Q2	
Figshare, PLoS.figshare, SpringerNature.figshare.com	658	54,96%	1183	33,45%	133	51,55%
ResearchGate	231	16,75%	877	24,80%	47	14,20%
Catalog.data.gov	23	1,66%	64	1,80%	12	4,65%
Statista	49	3,55%	194	5,48%	12	4,65%
Omicsdi	126	0,58%	294	8,31%	12	4,65%
Datamed	44	3,19%	133	3,76%	1	0,38%
Search.datacite	42	3,04%	225	6,36%	1	0,38%
Otros proveedores de datos	106	7,60%	566	16,00%	44	43,50%
Número total de datasets recuperados	1.379	100%	3.536	100%	258	100%

Como se ha comentado en el análisis de los resultados, un 37% de los conjuntos de datos científicos no son reutilizables o accesibles. Ello también es debido en parte a que proveedores como *ResearchGate* y *Statista* en muchas ocasiones limitan los datos mostrados, requiriendo registro, suscripción o pago previo para un acceso completo. Este tipo de prácticas están lejos del concepto de ciencia abierta y colaborativa, abriendo la reflexión sobre la cesión de derechos de los datasets en los principales proveedores y repositorios de datos.

> Siempre se produce un retraso en la publicación de datos con respecto a las fechas de los brotes o epidemias

Otro asunto relevante es la adecuación de los metadatos *Schema.org* a la descripción de los conjuntos de datos. Se ha observado que *Google Dataset Search* recopila la información sita en los distintos proveedores de datos. Cabe indicar que los metadatos *Schema.org* son un formato muy versátil, ya que adopta los principales asuntos de descripción para este tipo documental. Por ejemplo, la relación de pertenencia del dataset a otras colecciones de datos, resumen, autoría y su filiación, citas, comentarios, condiciones de acceso, estado de la edición, fechas de creación y actualización, ediciones, codificación, enlace del recurso, distribución, cobertura espacial y temporal, entre otros. Si bien resulta un modelo muy completo, si se cuida la descripción de cada apartado, existen dos aspectos fundamentales, que aún no han sido abordados adecuadamente:

- el control de las versiones de los datasets, con el fin de rastrear los cambios producidos y facilitar su recuperación. En este sentido, los metadatos *Schema.org* deberían incluir una *tupla* en la que se asocien las fechas de las versiones del dataset con su enlace al archivo original y la mención de responsabilidad, similar a la operativa ya conocida en repositorios de software como *GitHub* (**Blischak**; **Davenport**; **Wilson**, 2016);
- la falta de una definición de las estructuras de datos del dataset, lo que dificulta la unificación de conjuntos de datos con propósitos y temáticas similares. Esto es, la introducción de un campo que defina a modo de lista de valores separados por comas, la cabecera de campos del dataset en cuestión. Esto facilita la identificación y comparación rápida de la colección de datasets, a fin de descubrir cuáles pueden agruparse para generar grandes colecciones de datos. Por ende, también simplifica la automatización del mapeado de correspondencias entre campos de conjuntos de datos diferentes para su fusión, si fuera necesario.

La normalización de los metadatos busca garantizar la coherencia en el formato de los datasets recuperados por *Google Dataset Search* y poder ofrecer a los usuarios una experiencia de búsqueda significativa y unificada (**Canino**, 2019). Por ejemplo, que las actualizaciones de la situación epidemiológica del Covid-19 no se encuentren disgregadas en provincias y fechas de actualización, sino que puedan estudiarse en un único dataset que las integre todas, manteniendo la identificación de su procedencia, fecha y versión. Este tipo de casos es frecuente entre los resultados del buscador.

En cuanto a la reutilización de los datasets, cabe afirmar que su difusión y uso es su principal finalidad, ya que en la mayoría de los casos supone el registro de las pruebas, experimentos y observaciones científicas. Esto concede al dataset el valor de prueba científica en la que se basan las conclusiones y resultados de muchos artículos científicos. La naturaleza de este tipo documental lo convierte en un recurso o fuente de información valiosa, siempre y cuando sea posible su correcta identificación y agregación. En esta tarea se halla la minería de datos y textos, y en concreto las técnicas de *scraping* (**Singhal**; **Srivastava**, 2013), que hacen posible la extracción automática de datasets procedentes de distintos proveedores de datos, o bien, como es el caso, de un buscador especializado. Ello permite centrar la búsqueda en una temática o asunto concreto para obtener los conjuntos de datos más relevantes y procesarlos en una o varias tablas de una base de datos. Obviamente, aún no existe un método completamente automatizado por el cual pueda mapearse o corresponderse automáticamente todos los campos de un dataset. De aquí la importancia de contar con un sistema de metadatos adecuado que supla este problema, como se viene explicando.

Sucesivamente, la información procedente de los datasets puede ser enriquecida por otras fuentes de información hasta componer una colección más completa con la que obtener un nuevo análisis científico, en lo que se viene denominando proceso de curación de los datos científicos (**Karasti**; **Baker**; **Halkola**, 2006). La variedad de los datasets en torno a un objeto de estudio, así como su volumen, favorecen su reutilización en el análisis de *big data*. Esto es así porque la finalidad es la correlación de los conjuntos de datos procedentes de diversas fuentes. Por ejemplo, la relación entre los compuestos de los medicamentos que se emplean en el tratamiento de la Covid-19 y las evoluciones de los pacientes, los historiales clínicos y sus pruebas hematológicas (**Wang**; **Ng**; **Brook**, 2020). Constituyen conjuntos de datos diferenciados que son indexados, agrupados y relacionados, para inferir una relación de causa y efecto o bien una distribución de probabilidad acumulada de Pareto que facilite el diagnóstico del paciente (**Ahlawat**; **Chug**; **Singh**, 2019). Esto implica el almacenamiento multidimensional de los conjuntos de datos (**Elmeiligy**; **El-Desouky**; **Elghamrawy**, 2020), en donde cada dimensión es un factor de análisis, que es descompuesto para su identificación y clasificación en nodos, también denominados conjuntos de pares, para ser correlacionados con otros nodos procedentes de otros datasets. Esto es posible gracias a las técnicas de *Map-Reduce*, capaces de cuantificar la frecuencia de aparición de los elementos de cada nodo, dando como resultado un valor combinado y ordenado que refleja el peso de cada relación (**Khashan** *et al.*, 2020). Este tipo de resultados ayuda a determinar qué factores son decisivos en la mejoría o el agravamiento de una enfermedad, propiciando el posterior desarrollo de los modelos de aprendizaje automático.

> Sólo el 52% de los datos son científicos y sólo el 15% reutilizables

4. Conclusiones

1. En este trabajo se demuestra que no todos los resultados mostrados por el buscador son datasets. Si bien hay confusión con el término, se debe incidir en que un dataset o conjunto de datos es aquella colección de datos estructurada y delimitada en sus valores, de forma que puede ser reaprovechada en bases de datos, hojas de cálculo, programas de análisis estadístico y *big data*. Son formatos propios de los datasets el CSV de valores separados por comas, el SQL por constituir el lenguaje de consulta estructurado de bases de datos, y los derivados de XML por permitir la precisión de valores mediante etiquetas y estructuras extensibles de marcas. De hecho, según **Qian**, **Bailey** y **Leckie** (2006), un dataset es una colección de objetos o datos, representados de forma sucesiva siguiendo un patrón o esquema de tabulación que facilita su instanciación y recopilación.

> No todos los resultados mostrados por el buscador son datasets

2. El análisis demuestra que aproximadamente el 15% de los resultados son datasets reutilizables y de procedencia científica demostrada, lo que representa una cantidad reducida de la información disponible en el buscador. Si bien la ciencia abierta se está consolidando en el plano de los artículos científicos (**Mckiernan** *et al.*, 2016), no lo está haciendo en el campo de los conjuntos de datos. Esta conclusión puede deberse a la falta de supervisión, filtrado y evaluación de los datasets, previa indexación en el buscador. Según indica *Google*, la información de los conjuntos de datos se recopila directamente a través de los metadatos *Schema.org*. Quizá este procedimiento no desencadena una verdadera revisión del tipo de datos, formato, procedencia, calidad y fiabilidad de los mismos, resultando en un registro automático de la información. La plataforma de *Google* aún debe mejorar su proceso de entrada de datos y verificación, si su propósito es convertirse en una fuente fiable de datos científicos. Otros aspectos mejorables son:

- limitación de resultados a tan sólo un centenar por búsqueda,
- carencias del sistema de filtrado para distinguir correctamente los formatos de descarga según sus extensiones, la procedencia científica o comercial del dataset, país o región de recopilación, idioma, fechas de publicación y actualización, oficialidad de los datos, filiación, procedencia, sujetos productores secundarios o colaboradores, proveedor o agregador de datos, e intervalos de fechas.

3) *Google Dataset Search* se muestra como un agregador singular, que obtiene la información de múltiples proveedores de datos, entre los cuales se halla una desigualdad cuantitativa reseñable, al menos en los resultados obtenidos para familias de virus ARN. El principal proveedor de datos es *Figshare*. En la búsqueda general de familias de virus (Q1) recupera el 54,96% de información, en las búsquedas específicas de cada familia (Q2) el 33,45% y en la muestra específica de Q2 el 51,55%. En cambio, otros proveedores como *Catalog.data.gov*, *Statista*, *Omicsdi*, *Datamed* y *Search.datacite* no superan el 9% en cada una de las tres búsquedas realizadas.

4) Con el fin de aumentar el porcentaje de datasets reutilizables, *Google* debería recoger entre sus criterios de selección, cribado o filtrado de los datasets cuyos formatos sean CSV, XLS, SQL y XML y establecer una diferenciación clara entre conjuntos de datos comerciales, científicos, artículos científicos y otros productos de las publicaciones científicas, tales como figuras, ilustraciones y presentaciones. Se ha constatado que hay una gran diferencia entre los datos que ofrece el buscador en comparación con los datos reales. De hecho, sólo el 52% de los datasets son de procedencia científica, siendo realmente reutilizables el 15%. Un 37% de los resultados fueron mal catalogados, al considerarse datasets las ilustraciones, figuras y presentaciones de las publicaciones científicas.

5) La investigación parece confirmar retos que aún deben ser superados, tanto por los proveedores de datos como por los buscadores de datasets. En concreto:

- La normalización y perfeccionamiento del formato de metadatos para el intercambio de conjuntos de datos, a fin de que puedan distinguirse las versiones y adiciones de los autores e instituciones participantes en la edición.
- La indexación a texto completo de los datasets, incluyendo la relación de campos, cifras y cadenas de caracteres registradas, de forma que la recuperación no dependa exclusivamente de la descripción aportada en los metadatos.
- Mostrar las relaciones entre los artículos científicos y los datasets en los que se basan, para así poder estudiar su impacto en los avances científicos.
- Clasificar los conjuntos de datos según su disciplina científica, aplicaciones, cobertura temporal y geográfica.
- Desarrollar técnicas de *big data* para la detección de patrones de similaridad y correlación entre datasets, con objeto de facilitar al investigador la selección de los conjuntos de datos adecuados.

6) Se ha observado un notable incremento en la publicación de datasets de virus a partir de 2016, que coincide temporalmente con la epidemia de Ébola y Zika. Entre los virus de ARN con más datasets destacan el virus de la Coriomeningitis, el virus Nipah, el Rhinovirus, el SARS-CoV y recientemente de forma muy destacada el SARS-CoV-2. De hecho, han superado la ratio de publicación de datasets en un período de apenas 3 meses. Ninguno de los virus analizados alcanza estos niveles, lo cual demuestra una reacción evidente de la comunidad científica. Así pues, puede afirmarse que la mortalidad y morbilidad de un virus en la población son factores que están intrínsecamente vinculados

> Si bien la ciencia abierta se está consolidando en el plano de los artículos científicos, no lo está haciendo en el campo de los conjuntos de datos

con el número de datasets y *papers* publicados. Esto se demuestra al cotejar las cronologías de las principales epidemias, siendo especialmente cierto en el caso del H1N1, Ébola, Zika y SARS-CoV-2. También se puede concluir que la gravedad de una pandemia, la transparencia en las investigaciones, así como las dificultades para recopilar datos y diseñar métodos científicos para la elaboración de ensayos clínicos, pueden ser algunas de las causas de una baja frecuencia de publicación de los conjuntos de datos, sobre todo si se compara con otros tipos documentales, como los *papers* científicos. Este desequilibrio se amplía cuando se espera que el conjunto de datos sea de acceso abierto en su compleción, siendo una barrera más que deberá superar el esquema de ciencia abierta.

> *Google* debe mejorar su proceso de entrada de datos y verificación, si su propósito es convertirse en una fuente fiable de datos científicos

5. Futuras líneas de investigación

En este trabajo se demuestra la importancia de los datasets en el contexto de la Información y Documentación, ya que son una fuente de conocimiento fundamental para la producción científica. Las colecciones de datos son el tipo documental primario que recoge las observaciones de un experimento científico, o bien sus indicadores y factores de evaluación. Por tanto, con carácter general cualquier investigación que aborde su mejor conocimiento, administración, recuperación y procesamiento, facilitará el avance y difusión de la Ciencia. No debe olvidarse que la Documentación, como ciencia auxiliar, requiere un conocimiento abierto de las necesidades de información y de los tipos documentales que se demandan y emplean. Por tanto, caben futuras investigaciones relacionadas, por ejemplo:

- Análisis de los datasets de virus ADN y su comparación con los resultados de virus ARN. Replicando el mismo método de investigación, pueden tomarse como base de consulta los virus de tipo ADN, con la finalidad de ser comparados con los resultados de esta investigación. De esta forma sería posible confirmar si este otro tipo de virus tiene una ratio de publicación diferente, así como su incidencia en la publicación de artículos científicos, su dependencia de factores externos como epidemias o brotes, o bien la presencia de elementos diferenciadores con respecto a otro conjunto de datos. Los resultados obtenidos pueden proporcionar la experiencia necesaria para el diseño de mejores aplicaciones de búsqueda y agregación para este tipo de conjuntos de datos en el área de conocimiento de Virología.
- Replicación del método de investigación de datasets en otras áreas de conocimiento o temáticas. El método empleado es aplicable a cualquier tema. Esto es así ya que puede variarse la sección del tesauro de consulta o ser sustituido por un vocabulario controlado, normalizado y reconocido por la comunidad científica. Esto permite consultas dirigidas, conforme a la terminología oficial, en el buscador de datasets y obtener resultados homologables. Este tipo de estudios permitiría conocer qué áreas de conocimiento generan más colecciones de datos, por qué motivos, en qué condiciones, así como su correlación con las principales publicaciones científicas.
- Estudio de la correlación entre publicaciones científicas, el uso de conjuntos de datos y su valor. Puede afirmarse que las investigaciones científicas apoyadas por datasets, provenientes de la experimentación, observación y recopilación de datos, se encuentran en mejores condiciones para la demostración y justificación empírica que aquellas que no gozan de dichos fundamentos. Si bien esta hipótesis resulta razonable, cabría esperar su confirmación mediante el análisis de una muestra de artículos, basados en datasets científicos y de acceso abierto, en la que se analizaría su peso en la obtención de citas. Esto es, determinar el grado de influencia que este aspecto supone, para el éxito de un artículo científico.
- Diseño de un formato de metadatos adecuado a los datasets científicos. Si bien el formato de metadatos *Schema.org* es el referente actual, parece lógico plantear una mejora o actualización de sus campos, para que se adapte al control de versiones y la identificación de las estructuras de datos de los datasets para favorecer su agregación. Esta investigación, ayudaría a mejorar la recuperación de este tipo documental en buscadores como *Google Dataset Search*, brindando la oportunidad de simplificar la tarea del investigador.

> Se demuestra la importancia de los datasets en el contexto de la Información y Documentación, ya que constituyen una fuente de conocimiento fundamental para la producción científica

Nota del editor

1. Aunque "dataset" se puede traducir al castellano como "conjunto de datos", hemos mantenido la palabra inglesa por ser más corta y por tener un significado más preciso del concepto y su contexto.

6. Referencias

Ahlawat, Khyati; **Chug, Anuradha**; **Singh, Amit-Prakash** (2019). "Empirical evaluation of Map Reduce based hybrid approach for problem of imbalanced classification in big data". *International journal of grid and high performance computing*, v. 11, n. 3, pp. 23-45.
https://doi.org/10.4018/IJGHPC.2019070102

Bekelman, Justin E.; **MPhil, Yan-Li**; **Gross, Cary P.** (2003). "Scope and impact of financial conflicts of interest in biomedical research: a systematic review". *Jama*, v. 289, n. 4, pp. 454-465.
https://doi.org/10.1001/jama.289.4.454

Blischak, John D.; **Davenport, Emily R.**; **Wilson, Greg** (2016). "A quick introduction to version control with Git and GitHub". *PLoS computational biology*, v. 12, n. 1.
https://doi.org/10.1371/journal.pcbi.1004668

Brickley, Dan; **Burgess, Matthew**; **Noy, Natasha** (2019). "Google Dataset Search: Building a search engine for datasets in an open web ecosystem". In: *Proceedings of the 19th World wide web conference* (*WWW'19*), pp. 1365-1375.
https://doi.org/10.1145/3308558.3313685

Broder, Andrei (2002). "A taxonomy of web search". *ACM Sigir forum*, v. 36, n. 2, pp. 3-10.
https://doi.org/10.1145/792550.792552

Canino, Adrienne (2019). "Deconstructing Google Dataset Search". *Public services quarterly*, v. 15, n. 3, pp. 248-255.
https://doi.org/10.1080/15228959.2019.1621793

Chen, Emily; **Lerman, Kristina**; **Ferrara, Emilio** (2020). "Tracking social media discourse about the Covid-19 pandemic: Development of a public coronavirus Twitter data set". *JMIR public health and surveillance*, v. 6, n. 2.
https://arxiv.org/abs/2003.07372
https://doi.org/10.2196/19273

Chen, Serena H.; **Young, M. Todd**; **Gounley, John**; **Stanley, Christopher**; **Bhowmik, Debsindhu** (2020). "Distinct structural flexibility within SARS-CoV-2 spike protein reveals potential therapeutic targets". *BioRxiv*.
https://doi.org/10.1101/2020.04.17.047548

Corrales-Garay, Diego; **Ortiz-de-Urbina-Criado, Marta**; **Mora-Valentín, Eva-María** (2019). "Knowledge areas, themes and future research on open data: A co-word analysis". *Government information quarterly*, v. 36, n. 1, pp. 77-87.
https://doi.org/10.1016/j.giq.2018.10.008

Dick, George W. A.; **Kitchen, Stuart F.**; **Haddow, Alexander J.** (1952). "Zika virus (I). Isolations and serological specificity". *Transactions of the Royal Society of Tropical Medicine and Hygiene*, v. 46, n. 5, pp. 509-520.
https://doi.org/10.1016/0035-9203(52)90042-4

Elmeiligy, Manar A.; **El-Desouky, Ali I.**; **Elghamrawy, Sally M.** (2020). "A multi-dimensional big data storing system for generated Covid-19 large-scale data using Apache Spark". *arXiv preprint.*
https://arxiv.org/abs/2005.05036

Emond, Ronald T.; **Evans, Barry**; **Bowen, Ernest-Thomas**; **Lloyd, Graham** (1977). "A case of Ebola virus infection". *British medical journal*, v. 2, n. 6086, pp. 541-544.
https://doi.org/10.1136/bmj.2.6086.541

Google Search (2020). *Dataset*.
https://developers.google.com/search/docs/data-types/dataset

Haleem, Abid; **Javaid, Mohd**; **Khan, Ibrahim-Haleem**; **Vaishya, Raju** (2020). "Significant applications of big data in Covid-19 pandemic". *Indian journal of orthopaedics*, v. 54, n. 7.
https://doi.org/10.1007/s43465-020-00129-z

Hawking, David; **Craswell, Nick**; **Bailey, Peter**; **Griffihs, Kathleen** (2001). "Measuring search engine quality". *Information retrieval*, v. 4, n. 1, pp. 33-59.
https://doi.org/10.1023/A:1011468107287

Hawking, David; **Craswell, Nick**; **Thistlewaite, Paul**; **Harman, Dona** (1999). "Results and challenges in web search evaluation". *Computer networks*, v. 31, n. 11-16, pp. 1321-1330.
https://doi.org/10.1016/S1389-1286(99)00024-9

Hernández-Pérez, Tony (2016). "En la era de la web de los datos: primero datos abiertos, después datos masivos". *El profesional de la información*, v. 25, n. 4, pp. 517-525.
https://doi.org/10.3145/epi.2016.jul.01

Howe, Nicola; **Giles, Emma**; **Newbury-Birch, Dorothy**; **McColl, Elaine** (2018). "Systematic review of participants' attitudes towards data sharing: a thematic synthesis". *Journal of health services research & policy*, v. 23, n. 2, pp. 123-133.
https://doi.org/10.1177/1355819617751555

Irwin, Richard S. (2009). "The role of conflict of interest in reporting of scientific information". *Chest*, v. 136, n. 1, pp. 253-259.
https://doi.org/10.1378/chest.09-0890

Johansson, Michael A.; **Saderi, Daniela** (2020). "Open peer-review platform for Covid-19 preprints". *Nature*, v. 579, n. 7797.
https://doi.org/10.1038/d41586-020-00613-4

Karasti, Helena; **Baker, Karen S.**; **Halkola, Eija** (2006). "Enriching the notion of data curation in e-science: data managing and information infrastructuring in the long term ecological research (LTER) network". *Computer supported cooperative work*, v. 15, n. 4, pp. 321-358.
https://doi.org/10.1007/s10606-006-9023-2

Khashan, Eman A.; **El-Desouky, Ali I.**; **Fadel, Magdy**; **Elghamrawy, Sally M.** (2020). "A big data based framework for executing complex query over Covid-19 datasets (Covid-QF)". *arXiv preprint arXiv:2005.12271*.
https://arxiv.org/abs/2005.12271

King, John-Douglas; **Li, Yuefeng**; **Tao, Xiaohui**; **Nayak, Richi** (2007). "Mining world knowledge for analysis of search engine content". *Web intelligence and agent systems: An international journal*, v. 5, n. 3, pp. 233-253.
https://dl.acm.org/doi/10.5555/1377776.1377777

Landau, Yuval; **Kiryati, Nahum** (2019). "Dataset growth in medical image analysis research". *Arxiv.org*.
https://arxiv.org/abs/1908.07765

Le-Guillou, Ian (2020). "Covid-19: How unprecedented data sharing has led to faster-than-ever outbreak research". *Horizon. The UE research & innovation magazine*, 23 March.
https://horizon-magazine.eu/article/covid-19-how-unprecedented-data-sharing-has-led-faster-ever-outbreak-research.html

Lewandowski, Dirk (2015). "Evaluating the retrieval effectiveness of web search engines using a representative query sample". *Journal of the Association for Information Science and Technology*, v. 66, n. 9, pp. 1763-1775.
https://asistdl.onlinelibrary.wiley.com/doi/10.1002/asi.23304
https://doi.org/10.1002/asi.23304

López-Borrull, Alexandre; **Ollé-Castellà, Candela**; **García-Grimau, Francesc**; **Abadal, Ernest** (2020). "Plan S y ecosistema de revistas españolas de ciencias sociales hacia el acceso abierto: amenazas y oportunidades". *El profesional de la información*, v. 29, n. 2.
https://doi.org/10.3145/epi.2020.mar.14

Marcial, Laura-Haak; **Hemminger, Bradley M.** (2010). "Scientific data repositories on the Web: An initial survey". *Journal of the American Society for Information Science and Technology*, v. 61, n. 10, pp. 2029-2048.
https://doi.org/10.1002/asi.21339

McKiernan, Erin C.; **Bourne, Philip E.**; **Brown, C. Titus**; **Buck, Stuart**; **Kenall, Amye**; **Lin, Jennifer**; **McDougall, Damon**; **Nosek, Brian A.**; **Ram, Karthik**; **Soderberg, Courtney K.**; **Spies, Jeffrey R.**; **Thaney, Kaitlin**; **Updegrove, Andrew**; **Woo, Kara H.**; **Yarkoni, Tal** (2016). "Point of view: How open science helps researchers succeed". *Elife*, v. 5, e16800.
https://doi.org/10.7554/eLife.16800.001

Mello, Michelle M.; **Lieou, Van**; **Goodman, Steven N.** (2018). "Clinical trial participants' views of the risks and benefits of data sharing". *New England journal of medicine*, v. 378, n. 23, pp. 2202-2211.
https://doi.org/10.1056/NEJMsa1713258

Nosek, Brian A.; **Alter, George**; **Banks, George C.**; **Borsboom, Denny**; **Bowman, Sara D.**; **Breckler, Steven J.**; **Buck, Stuart**; **Chambers, Christopher D.**; **Chin, Gilbert**; **Christensen, Garret**; **Contestabile, M.**; **Dafoe, A.**; **Eich, Eric**; **Freese, J.**; **Glennerster, R.**; **Goroff, D.**; **Green, Donald P.**; **Hesse, Bradford W.**; **Humphreys, M.**; **Ishiyama, John**; **Karlan, D.**; **Kraut, A.**; **Lupia, A.**; **Mabry, Patricia L.**; **Madon, T.**; **Malhotra, N.**; **Mayo-Wilson, Evan**; **McNutt, M.**; **Miguel, Edward**; **Levy-Paluch, Elizabeth**; **Simonsohn, U.**; **Soderberg, Courtney**; **Spellman, Barbara A.**; **Turitto, J.**; **VandenBos, Gary-Roger**; **Vazire, Simine**; **Wagenmakers, E. J.**; **Wilson, R.**; **Yarkoni, T.** (2015). "Promoting an open research culture". *Science*, v. 348, n. 6242, pp. 1422–1425.
https://doi.org/10.1126/science.aab2374

Polonetsky, Jules; **Tene, Omer**; **Finch, Kelsey** (2016). "Shades of gray: Seeing the full spectrum of practical data de-intentification". *Santa Clara law review.* v. 56, n. 593, pp. 593-618.
https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=2827&context=lawreview

Qian, Xiaoyuan; **Bailey, James**; **Leckie, Christopher** (2006). "Mining generalised emerging patterns". In: Sattar, Abdul; Kang, Byeong-Ho (eds.). *Australasian joint conference on artificial intelligence*. Berlin, Heidelberg: Springer, pp. 295-304. ISBN: 978 3 540 49788 2
https://doi.org/10.1007/11941439_33

Saheb, Tahereh; **Izadi, Leila** (2019). "Paradigm of IoT big data analytics in healthcare industry: a review of scientific literature and mapping of research trends". *Telematics and informatics*, v. 41, pp. 70-85
https://doi.org/10.1016/j.tele.2019.03.005

Schneier, Bruce (2012). "Securing medical research: A cybersecurity point of view". *Science*, v. 336, n. 6088, pp. 1527-1529.
https://doi.org/10.1126/science.1224321

Science Europe (2019). *Plan S: Making full and immediate Open Access a reality.*
https://www.scienceeurope.org/coalition-s

Singhal, Ayush; **Srivastava, Jaideep** (2013). "Data extract: Mining context from the web for dataset extraction". *International journal of machine learning and computing,* v. 3, n. 2, pp. 219-223.
https://doi.org/10.7763/IJMLC.2013.V3.306

Wang, C. Jason; **Ng, Chun Y.**; **Brook, Robert H.** (2020). "Response to Covid-19 in Taiwan: big data analytics, new technology, and proactive testing". *Jama*, v. 323, n. 14, pp. 1341-1342.
https://doi.org/10.1001/jama.2020.3151

Weston, Sara J.; **Ritchie, Stuart J.**; **Rohrer, Julia M.**; **Przybylski, Andrew K.** (2019). "Recommendations for increasing the transparency of analysis of preexisting data sets". A*dvances in methods and practices in psychological science,* v. 2, n.3, pp. 214-227.
https://doi.org/10.1177/2515245919848684

Zhou, Chenghu; **Su, Fenzhen**; **Pei, Tao**; **Zhang, An**; **Du, Yunyan**; **Luo, Bin**; **Cao, Zhidong**; **Wang, Juanle**; **Yuan, Wen**; **Zhu, Yunqiang**; **Song, Ci**; **Chen, Jie**; **Xu, Jun**; **Li, Fujia**; **Ma, Ting**; **Jiang, Lili**; **Yan, Fengqin**; **Yi, Jiawei**; **Hu, Yunfeng**; **Liao, Yilan**; **Xiao, Han** (2020). "Covid-19: challenges to GIS with big data". *Geography and sustainability*, v. 1, n, 1, pp. 77-87.
https://doi.org/10.1016/j.geosus.2020.03.005