uc3m | Universidad Carlos III de Madrid

Máster Universitario en Bibliotecas, Archivos
y Continuidad Digital (2019-2021)

Trabajo Fin de Máster

Análisis del estado de publicación de los datos de investigación en artículos españoles sobre COVID-19: una primera aproximación

---

Roxana Paulina Cerda Cosme

Tutora

Prof. Dra. Eva Méndez

Madrid, 22 de Febrero 2021

DETECCIÓN DEL PLAGIO

La Universidad utiliza el programa **Turnitin Feedback Studio** para comparar la originalidad del trabajo entregado por cada estudiante con millones de recursos electrónicos y detecta aquellas partes del texto copiadas y pegadas. Copiar o plagiar en un TFM es considerado una **Falta Grave**, y puede conllevar la expulsión definitiva de la Universidad.

Índice de contenidos

Índice de contenidos 2

Índice de Tablas 3

Índice de Figuras 3

1. Introducción 4

1.1. Motivación y estructura del trabajo 4

1.2. Preguntas de investigación y objetivos 5

1.3. Metodología 7

1.3.1. Definición y alcance del objeto de estudio 7

1.3.2. Materiales y fuentes 7

1.3.3. Métodos y limitaciones de análisis 9

2. Estado de la cuestión: los datos de investigación relacionados con COVID-19 11

2.1. Marco teórico general 11

2.1.1. Datos de investigación: Definición, tipología y ciclo de vida de los datos 11

2.1.2. Aspectos técnicos de la apertura de datos de investigación 17

2.1.3. Infraestructuras y datos de investigación 25

2.2. Marco regulador para la gestión y compartición de datos de investigación 27

2.2.1. Directiva de la Unión Europea 2019/1024, de 20 de junio de 2019 27

2.2.2. Programa Horizonte 2020 (2014-2020) 28

2.2.3. Programa Horizonte Europa 2021-2027 29

2.2.4. Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación 29

2.2.5. La Estrategia Española de Ciencia, Tecnología e Innovación 2021-2027 30

2.3. COVID-19 y el cambio de paradigma en la comunicación científica 31

2.4. Iniciativas Open Science para publicaciones y datos sobre la COVID-19 36

2.4.1. *Manifesto for European Union COVID-19 Research* 36

2.4.2. Compartir datos de investigación y hallazgos relevantes para el nuevo brote de coronavirus (NCOV) 37

2.4.3. *Data Together COVID-19* 37

3. Análisis de los datos de investigación de las publicaciones españolas sobre COVID-19: Resultados y Discusión 38

4. Conclusiones 52

5. Bibliografía/Referencias 53

Índice de Tablas

Tabla 1: Atributos para la evaluación de datos depositados en repositorios. Fuente: Elaboración propia 11
Tabla 2: Licencias Creative Commons. Fuente: CreativeCommons. 24
Tabla 3: Tipos de repositorios. Fuente: SpringerNature 26
Tabla 4: Accesibilidad y vinculación de los artículos y sus datos. Fuente: Elaboración propia 41
Tabla 5: Tipos de datos abiertos y formatos de archivos. Fuente: Elaboración propia. 42
Tabla 6: Licencias aplicadas a los datos. Fuente: Elaboración propia. 43
Tabla 7: Agencias financiadoras y número de proyectos. Fuente: Elaboración propia . 44
Tabla 8: número de artículos publicados con datos asociados por editorial. Fuente: Elaboración propia 45
Tabla 9: Política de datos de investigación por editorial. Fuente: Elaboración propia.. 46
Tabla 10: Contenido de los datos de investigación. Fuente: Elaboración propia. 48
Tabla 11: Datos estadísticos por repositorio. Fuente: Elaboración propia 49
Tabla 12: Contenido de los conjuntos de datos finales publicados de la muestra inicial. Fuente: Elaboración propia. 51

Índice de Figuras

Figura 1: Diagrama de Flujo PRISMA. Fuente: Moher et al., 2009 9
Figura 2: Modelo de ciclo de vida de datos científicos de USGS. Fuente: Faundeen et al., 2014 15
Figura 3: DataONE: Ciclo de vida de los datos. Fuente: Strasser, 2012 15
Figura 4: Ciclo de vida de conservación de DCC. Fuente: Higgins, 2008. 17
Figura 5: Principios FAIR. Fuente: Australian National Data Service (ANDS) 19
Figura 6: Iniciativas GOFAIR. Fuente: GOFAIR 20
Figura 7: Herramientas y software desarrolladas por FAIRSFAIR. Fuente: FAIRSFAIR... 21
Figura 8: Preprints publicados sobre COVID-19 o SARS-CoV-2 en repositorios temáticos. Fuente: Kwon, D. (2020) 32
Figura 9: Datos de investigación vinculados a artículos españoles sobre la COVID-19. Fuente: Elaboración propia. 39
Figura 10: Datos depositados en repositorios especializados y generales. Fuente: Elaboración propia 40
Figura 11: Datos disponibles a solicitud del autor. Fuente: Elaboración propia..... 50

1. Introducción

La ciencia abierta es un movimiento que "afecta a la forma en que se lleva a cabo el trabajo científico y se caracteriza por un amplio esfuerzo de colaboración en cada etapa del proceso de investigación" (Gagliardi et al., 2015, p.1). Se presenta como "un cambio sistémico que permite mejorar la ciencia a través de formas abiertas y colaborativas para producir, compartir conocimientos y datos, lo antes posible en el proceso de investigación" (Comisión Europea, 2019, p.1). Este cambio es entendido como una transición de un paradigma a otro como respuesta a las nuevas necesidades generadas en un contexto determinado, "el cambio no está en lo que se hace, sino en cómo se hace" (Anglada & Abadal, 2018, p.293).

En el contexto actual, vivimos una pandemia mundial debido a la COVID-19 causada por el coronavirus SARS-CoV-2 que ha acelerado un cambio dentro del proceso de investigación y comunicación científica apoyados en la Ciencia Abierta. La pandemia "pone de manifiesto dos ineficiencias en el sistema de investigación: el abandono de la ciencia cerrada y el énfasis excesivo en las publicaciones de élite" (Larivière et al., 2020) evidenciando que el paradigma clásico de comunicación científica carece de soluciones eficientes para remediar las consecuencias del virus.

Las editoriales plantearon iniciativas colaborativas basadas en el acceso abierto a los artículos y los datos de investigación para "crear vasos comunicantes entre estructuras que antes no estaban comunicadas, precisamente por cuestiones de mercado, de competencia entre revistas" (Abellan et.al, 2020, p.254). Los datos de investigación toman una especial relevancia como el "resultado más valioso de muchos proyectos de investigación que se utilizan como fuentes primarias que sustentan la investigación científica y permiten la derivación de hallazgos teóricos o aplicados" (Bezjak et al., 2018) haciendo posible "ahorrar miles de millones de cada año y, al mismo tiempo, mejorar la salud y la longevidad de los ciudadanos y reducir la cantidad de daños a los que están expuestos" (Gøtzsche, 2012, p.236).

1.1. Motivación y estructura del trabajo

Estamos presenciando un cambio importante en la forma de hacer y compartir ciencia, que motiva la meta-investigación que ayude a mejorar la eficacia de la comunicación científica en situaciones de crisis como la pandemia. Las publicaciones (y los datos) se han incrementado en proporciones no registradas anteriormente frente a un virus y es esta característica la que nos lleva a preguntarnos cómo se está estructurando y validando esta información (datos y publicaciones) para que pueda ser encontrada, reutilizada, accesible e interoperable.

Por otra parte, estudiar el caso español es otra de las motivaciones del trabajo. España es el cuarto país con la tasa de retorno de inversión más alta de la Unión Europa dentro del Programa Horizonte 2020 (H2020) en el periodo 2014-2020. Se proyectó lograr un 9,5% para el 2020, pero según los resultados publicados por el Centro para el

Desarrollo Tecnológico Industrial (CDTI) para el año 2016 ya se había alcanzado el 9,4% y los dos últimos años se había logrado el 9,8% y 10,1% respectivamente; superando el objetivo proyectado al inicio del programa. Es el tercer país en términos de participación en proyectos de investigación ya que una de cada cuatro propuestas consiguió financiación de H2020 y cuenta con, al menos, un representante español. Además, es el primer país en liderazgo de proyectos de I+D+I en colaboración, al ser el coordinador del país en 940 de los proyectos financiados. Sin embargo, nos gustaría saber cuál es la respuesta española en la publicación sobre COVID-19 en lo que se refiere a la disponibilidad de los datos de investigación.

Este trabajo se centra en los datos de investigación que fueron la base para los artículos españoles publicados sobre la COVID-19 durante los primeros ocho meses del 2020. Se analizará la casuística de dichos datos, con la finalidad de identificar su disponibilidad en las plataformas de depósito y las condiciones en las que fueron depositados y/o compartidos. Como resultado de la investigación se espera tener una primera aproximación de referencia sobre qué datos, subyacentes a las publicaciones españolas sobre COVID, están publicados, compartidos y accesibles y, en su caso, en qué condiciones.

El trabajo se estructura en cuatro puntos o aspectos fundamentales. El primero de ellos, refleja todos los aspectos introductorios donde se desarrolla la motivación, preguntas de investigación y la metodología aplicada para el estudio. En el segundo apartado, realiza un necesario estado de la cuestión o estado del arte sobre los datos de investigación relacionados con la COVID-19 que incluye: el marco teórico y regulador, la situación actual de los datos de investigación, así como algunas alternativas o soluciones que se han planteado para compartir los datos de investigación. El tercer punto se centra en presentar y analizar los resultados de la investigación, haciendo hincapié en la discusión sobre los mismos. Finalmente, en el cuarto apartado se exponen las conclusiones finales.

1.2. Preguntas de investigación y objetivos

La reutilización de los datos de investigación, que es uno de los aspectos fundamentales en el ámbito de la ciencia abierta, se convierte en un aspecto esencial y urgente en el ámbito de la pandemia, crucial para para fomentar la innovación y la creación de posibles soluciones, entendimiento y avance científico. Para que se lleve a cabo se necesita que los datos vayan asociados a los artículos en plataformas de depósito de acuerdo a la naturaleza de los datos. De esta afirmación se generan las siguientes preguntas de investigación: ¿los artículos de investigación publicados sobre la COVID-19 publican también los datos asociados a la investigación? y, en su caso, ¿cuáles son las plataformas de depósito que utilizan los investigadores para depositar/publicar dichos datos? Además, teniendo en cuenta que el tema del depósito y publicación de datos no está aún implantada en la práctica de la

comunicación científica actual, hay muchos conceptos que se confunden y por ello, también nos preguntamos si los autores y editores ¿reconocen la diferencia entre material complementario y datos asociados al artículo?

Asimismo, hay otras preguntas que nos hacemos en relación a los atributos y características de los datos:

- ¿El depósito de los datos es un requisito obligatorio por las editoriales para la publicación del artículo?
- ¿Cuáles son los formatos y plataformas utilizados por los investigadores para depositar sus datos?
- ¿Qué tipo de licencias emplean y cuáles son las condiciones de uso?
- ¿Los datos están referenciados hacia el artículo y a su vez los artículos tienen el enlace a los datos?
- ¿Los datos cuentan con un identificador único y persistente?
- ¿Se puede tener acceso a las estadísticas de uso y citación de los datos?

Y, en el caso de que los artículos que no tengan los datos vinculados a la publicación, también se nos ocurren diversas preguntas ¿cuáles son los motivos para no tener los datos disponibles? ¿están los datos depositados, pero tienen un periodo de embargo? ¿Se puede acceder mediante una solicitud al autor?, en su caso, ¿cuáles serían las condiciones de acceso y de qué dependen estas?

Esta variedad y prolijidad de preguntas daría lugar a varios trabajos de investigación, sin embargo, plantearlas aquí de manera conjunta nos ayuda a dirimir el objetivo principal de este estudio, así como los objetivos específicos de nuestro trabajo de investigación, concebido como trabajo fin de máster.

En este sentido, el **objetivo general** es: estudiar en qué medida los artículos científicos españoles sobre COVID-19 publicados en los primeros ocho meses (Enero-Agosto) de la pandemia, publican los datos de investigación asociados a los mismos.

Asimismo, una vez analizado el objetivo general, nos gustaría llegar un poco más lejos, con los siguientes **objetivos específicos**: a) Analizar la casuística de los datos de investigación asociados a los artículos españoles sobre la COVID-19: tipos de datos y nivel de apertura y repositorios específicos de depósito y/o publicación, y b) Identificar, en la medida de lo posible, las características y atributos de dichos datos de investigación: si llevan asociados metadatos, si están identificados por identificadores persistentes, llevan asociadas licencias, y otros aspectos que pudieran ayudarnos a cualificar su nivel de coherencia en la publicación de datos.

1.3. Metodología

1.3.1. Definición y alcance del objeto de estudio

El objeto de estudio es, como se ha anticipado anteriormente, las publicaciones españolas sobre COVID-19. Para seleccionar dichas publicaciones y definir cuáles son objeto de nuestro estudio, se analizaron los artículos publicados en revistas indizadas en Scopus (Elsevier, 2020) entre enero y agosto de 2020. Se delimitó este periodo de tiempo, porque durante los primeros ocho meses del año sucedieron dos hechos importantes que nos ayudan a delimitar el espectro y alcance temporal de la muestra: por un lado, la declaración de emergencia internacional por parte de la Organización Mundial de la Salud (OMS) y, la definición temporal de la "primera ola" del virus en los países europeos que desató asimismo el ritmo de crecimiento de las publicaciones científicas a nivel global, llegando a 500 publicaciones diarias y 1.000 publicaciones semanales (Torres-Salinas, 2020).

1.3.2. Materiales y fuentes

Como material básico para la selección de la muestra de nuestro trabajo se eligió la base de datos Scopus porque combina "las características de PubMed y Web of Science. Estas características combinadas permiten una mayor utilidad, tanto para la investigación de la literatura médica como para las necesidades académicas (análisis de citas)" (Falagas et al., 2007, p.339). Estas dos características de la base de datos Scopus permitieron definir el alcance de nuestra investigación, ya que se utilizó el número de citas de los artículos, como uno de los filtros para seleccionar y construir la muestra final. También es relevante el hecho de que esta base de datos recoja características de PubMed porque los artículos analizados están relacionados con medicina en un porcentaje alto.

Para construir la **estrategia de búsqueda** se tomó como referencia los términos basados en los *Medical Subject Headings (MeSH)* creado por la *National Library of Medicine (NLM)*. Se utilizó la búsqueda avanzada, unificando los campos *TITLE + ABSTRACT + KEYWORDS* en un solo campo. Se aplicó el operador booleano OR para incluir cualquiera de los términos propuestos y el operador de proximidad W/n para delimitar el número máximo de términos entre las palabras claves:

```
TITLE-ABS-KEY ( "2019-nCoV" OR "COVID-19" OR "SARS-CoV-2" OR
"HCoV-2019" OR "hcov" OR "NCOVID-19" OR "severe acute
respiratory syndrome coronavirus 2" OR "severe acute respiratory
syndrome corona virus 2" OR "SARS-CoV2" OR covid2019 OR "COVID-
19" OR covid19 OR 2019ncov OR "2019 ncov" OR "novel coronaviru*"
OR "novel coronaviruses" OR "novel corona virus" OR "novel
corona*" OR covid19 OR "covid 19" OR "sars cov 2" OR sars2 OR
"new corona*" OR "new coronavirus" OR "coronavirus disease 2019"
OR "coronavirus infection" OR "COVID-19 illness" ) OR TITLE-ABS-
KEY ( ( coronaviru* OR "corona viru*" OR "pneumonia viru*" OR
cov OR ncov ) W/4 ( wuhan OR china OR novel ) )
```

Para limitar la búsqueda se aplicaron además cuatro filtros: “Artículo”, como tipo de documento, “España”, como país de afiliación del autor, “2020”, como año de publicación y se tomaron sólo los artículos cuyo estado de publicación era “final”.

```
LIMIT-TO ( DOCTYPE , "ar" )
LIMIT-TO ( AFFILCOUNTRY , "Spain" )
LIMIT-TO ( PUBYEAR , 2020 )
LIMIT-TO ( PUBSTAGE , "final" )
```

Así pues, la sentencia de búsqueda final fue:

```
TITLE-ABS-KEY ( "2019-nCoV" OR "COVID-19" OR "SARS-CoV-2" OR "HCoV-2019" OR "hcov" OR "NCOVID-19" OR "severe acute respiratory syndrome coronavirus 2" OR "severe acute respiratory syndrome corona virus 2" OR "SARS-CoV2" OR covid2019 OR "COVID-19" OR covid19 OR 2019ncov OR "2019 ncov" OR "novel coronaviru*" OR "novel coronaviruses" OR "novel corona virus" OR "novel corona*" OR covid19 OR "covid 19" OR "sars cov 2" OR sars2 OR "new corona*" OR "new coronavirus" OR "coronavirus disease 2019" OR "coronavirus infection" OR "COVID-19 illness" ) OR TITLE-ABS-KEY ( ( coronaviru* OR "corona viru*" OR "pneumonia viru*" OR cov OR ncov ) W/4 ( wuhan OR china OR novel ) ) AND ( LIMIT-TO ( AFFILCOUNTRY , "Spain" ) ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) ) AND ( LIMIT-TO ( DOCTYPE , "ar" ) ) AND ( LIMIT-TO ( PUBSTAGE , "final" ) )
```

La búsqueda se realizó el 31 de agosto del 2020 y se obtuvieron 666 resultados. Se exportó en formato de valores separados por comas (.CSV) de acuerdo a las categorías propias de Scopus. De este grupo inicial de resultados (666) se desestimaron: 350 artículos que no habían recibido ninguna cita (esto es con un número de citas =0); 3 cartas al editor; 1 artículo de opinión y 6 editoriales, porque no cumplían con los filtros antes mencionados. La muestra final para el análisis se constituyó finalmente con 306 artículos de investigación que representa el 45% de los resultados totales obtenidos, pero que implicaban una mayor representatividad para los objetivos del estudio (Fig. 2).

Para definir la sentencia de búsqueda o *query* final así como para filtrar los resultados, hemos utilizado el diagrama de flujo PRISMA (Fig. 1) donde se visualiza de manera esquemática la muestra final. La base de datos Scopus identifico 666 artículos que coincidían con la sentencia de búsqueda. De este grupo se eliminaron 350 artículos con un número de citas igual a cero quedando 316 artículos. Analizando los artículos se identificaron 10 que no eran artículos de investigación (6 editoriales, 1 artículo de opinión y 3 cartas al editor) que fueron excluidos de la muestra. De los 306 artículos restantes, se eliminaron 219 porque no tenían datos asociados. El número de artículos analizados e incluidos en la muestra final fue 87.

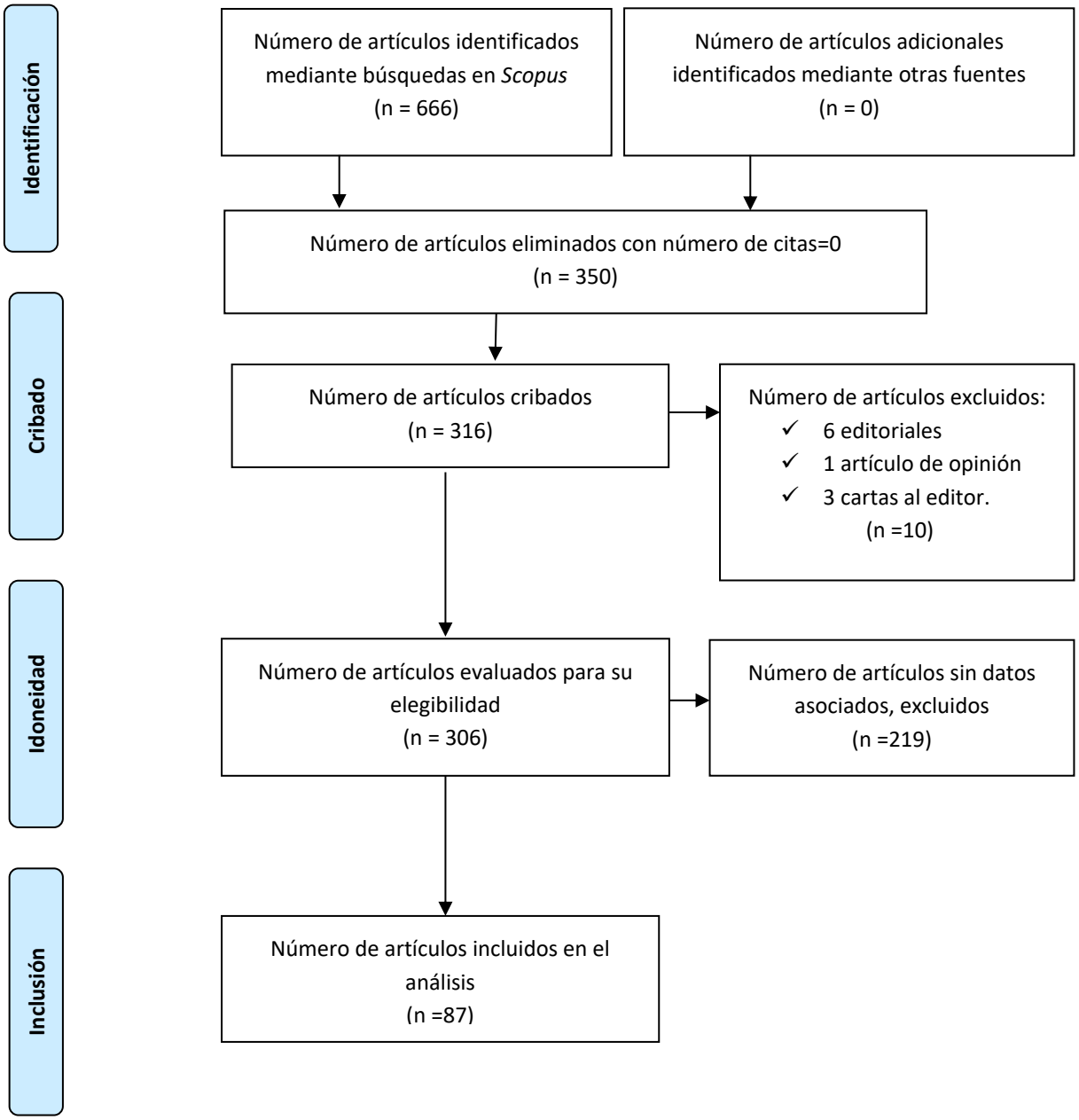

Figura 1: Diagrama de Flujo PRISMA. Fuente: Moher et al., 2009

1.3.3. Métodos y limitaciones de análisis

A la hora de establecer una metodología que nos permitiera estudiar de forma consistente los datos de investigación vinculados a las publicaciones científicas seleccionadas, nos encontramos con varias limitaciones, derivadas en general, de la falta de estándares y buenas prácticas consolidadas en la compartición de datos. La primera limitación se detectó al plantear contrastar los datos extraídos de Scopus, con los datos de Figshare y Zenodo. En el caso de Figshare, el mayor problema es que este repositorio multitemático no cuenta con la opción de descarga masiva. Para corroborar esta limitación se realizó la consulta el 09 de setiembre de 2020 al correo support@figshare.com, confirmando esta limitación en la respuesta del soporte del

repositorio: no se pueden realizar descargas masivas y para consultar los datos se debe abrir cada elemento por separado para descargar los archivos. En el caso de Zenodo, permite la exportación masiva de datos en diferentes formatos, pero al contrastar ambas tablas no se encontró una "llave" o dato externo que pueda vincular ambas tablas, porque los metadatos asignados al artículo, y los asignados al conjunto de datos, no son congruentes.

Estamos de acuerdo con la afirmación de Travieso Rodríguez & Ferreira Araújo (2019): "la disponibilidad de indicadores cuantitativos procedentes de bases de datos y recursos informativos no garantiza su comprensión global, a lo que se suman las particularidades entre disciplinas científicas y la dispersión de las fuentes" (p.6). Se deduce que los indicadores cuantitativos no son suficientes para evaluar los datos de investigación, es necesario establecer indicadores cualitativos que nos ayuden a conocer su naturaleza.

En este sentido, la segunda limitación que se detectó fue que la ausencia de indicadores de calidad e impacto en los escenarios de evaluación de datos de investigación como el nuestro. Las medidas o indicadores vinculados a los datos suelen estar relacionadas con la "apertura" o la disponibilidad de los datos, en lugar de medir la calidad, el impacto o la reutilización de los mismos (Konkiel, 2020). Esto nos lleva a la primera conclusión antes de comenzar el estudio: se evidencia un vacío dentro de la construcción de criterios normalizados para la evaluación de datos de investigación.

Así pues, el método principal que hemos adoptado en esta investigación es el estudio descriptivo y detallado basado en la casuística encontrada al revisar manualmente cada uno de los artículos que se ha adoptado en otros estudios (Vasilevsky et al., 2017; Travieso Rodríguez & Ferreira Araújo, 2019) y el conjunto de datos asociados a él.

El análisis se realizó en tres etapas que se describen a continuación:

1) En la primera etapa se analizaron 306 artículos y los metadatos de: autor, título, año de publicación, título de la revista, número de citas, DOI, enlace a artículo en Scopus, afiliación, editorial, idioma, tipo de documento, etapa de la publicación y tipo de acceso; con la finalidad de identificar aquellos que incluían datos asociados o material complementario → 87 artículos cumplieron con esta condición y 219 fueron excluidos.
2) En la segunda etapa, los 87 artículos se agruparon de acuerdo a las formas fundamentales que existen actualmente para compartir datos en el contexto científico: "i) agregar los datos brutos como material complementario a la publicación, ii) estar disponible a solicitud de los autores y iii) cargar los datos sin procesar en un repositorio de datos" (Sixto-Costoya et.al, 2020).
3) En la tercera etapa se aplicaron diferentes criterios de análisis para los grupos antes mencionados:

- Para los artículos con material complementario, se identificó y diferencio entre tablas y figuras (material complementario) y conjunto de datos (datos asociados).
- Para los artículos con datos asociados, pero que no están disponibles, se agruparon de acuerdo a las razones indicadas en la declaración de disponibilidad de datos.

- Para los artículos con datos depositados en repositorios, se tomó como referencia los atributos propuestos por (Assante et.al, 2016) y empleados en investigaciones anteriores (Travieso Rodríguez & Ferreira Araújo, 2019) para construir los nueve atributos (tabla 1) que se identificaron y analizaron de los datos de investigación.

Tabla 1: Atributos para la evaluación de datos depositados en repositorios. Fuente: Elaboración propia

Nº	Atributo	Descripción
1	Disponibilidad	Indicar el nombre del repositorio de datos donde se encuentran disponibles los datos.
2	Accesibilidad	Identificación de un URI que garantice el acceso a los datos.
3	Formato	Indicar el tipo de formato de los datos.
4	Licencia	Identificar el tipo de licencia y la descripción de las políticas que rigen la reutilización del conjunto de datos, incluidos los derechos de acceso.
5	Vinculación	Referenciar hacia el artículo y a su vez los artículos tienen el enlace a los datos.
6	Financiación	Indicar la agencia que subvencionó el proyecto.
7	Política editorial	Indicar las políticas editoriales sobre datos de investigación compartidos y asociados al artículo.
8	Contenido	Indicar que tipo de dato de investigación tiene asociado al artículo.
9	Estadísticas	Indicar los datos bibliométricos y/o altmétricos de los conjuntos de datos: visualizaciones, descargas, citas y/o métricas alternativas.

2. Estado de la cuestión: los datos de investigación relacionados con COVID-19

2.1. Marco teórico general

2.1.1. Datos de investigación: Definición, tipología y ciclo de vida de los datos

La Organización de Cooperación y Desarrollo Económicos (OCDE) definió formalmente el término “dato de investigación” en el año 2007 como:

> “Registros fácticos (puntuaciones numéricas, registros textuales, imágenes y sonidos) utilizados como fuentes primarias para la investigación científica, y

> que son comúnmente aceptados por la comunidad científica como necesarios para validar los resultados de la investigación. Un conjunto de datos de investigación constituye una representación sistemática y parcial del sujeto investigado" (p.13)

Como se establece en la base de datos Recolecta[1] "la naturaleza de los datos de investigación es muchísimo más variable y dependiente de la disciplina y de su particular ciclo de vida". Partiendo de este planteamiento se puede afirmar que la definición de dato de investigación depende del área del conocimiento donde se producen. Así, por ejemplo, en el área de la salud, la *National Institutes of Health (NIH)*[2] los define como el material fáctico registrado y aceptado en la comunidad científica según sea necesario para validar los resultados de la investigación. Para el área de las ciencias, la *National Science Foundation (NSF)*[3] afirma que son algo determinado por la comunidad de interés a través del proceso de revisión por pares y gestión de proyectos. Son esenciales no solo para investigadores individuales o equipos de investigación, también para la comunidad investigadora. En el área de Humanidades, la *National Endowment for the Humanities (NEH)*[4], hace una definición más inclusiva y considera datos de investigación a los materiales generados o recopilados durante el curso de la realización de la investigación.

La Comisión Europea[5] define los datos de investigación como información (en particular hechos o cifras) recopilada para ser examinada y servir como base para el razonamiento, la discusión o el cálculo.

No existe un consenso generalizado sobre la definición de los datos de investigación debido en parte, como apuntamos anteriormente, a que se conceptualizan de acuerdo al campo de investigación donde se generan, pero si podemos exhortar al acuerdo de que, independientemente de la disciplina científica donde se generen, los datos deben ser válidos, compartidos, heterogéneos y contextualizados dentro de las comunidades de investigación. Así pues, la tipología de los datos de investigación, responde normalmente al campo científico o disciplinar donde se generan y a la metodología. Los formatos no son los mismos cuando se investiga en el campo de la física, de la medicina, de las ciencias sociales o de las humanidades. Los formatos deben ser elegidos por la comunidad científica que genera los datos, y es susceptible de reutilizarlos. Por otro lado, los repositorios de datos deben estar preparados para

[1] https://recolecta.fecyt.es
[2] https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
[3] https://www.nsf.gov/data
[4] https://www.neh.gov/open
[5] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

almacenar estos formatos, sobre todo los repositorios generales o multidisciplinares, que permiten el depósito de datos de diferentes especialidades

El Consorcio de Servicios Universitarios de Cataluña [CSUC], (2020) determinó los tipos de datos de investigación según la fuente de la que provienen, distinguiendo entre:

- Observacionales: datos capturados en tiempo real (neuroimágenes, datos de muestras, datos de sensores, datos de encuestas, etc.).
- Experimentales: datos capturados en equipos de laboratorio (secuencias de genes, cromatogramas, datos de campos magnéticos, etc.).
- Simulación: datos generados a partir de modelos de prueba (climatológicas, matemáticas o modelos económicos, etc.).
- Derivados o compilados: datos reproducibles, pero de difícil reproducción (texto y minería de datos, modelos 3D, base de datos compilada, etc.).
- De referencia: conglomerado o conjunto de datos (bancos de datos de secuencias de genes, estructuras químicas, portales de datos espaciales, etc.).

La Comisión Económica para América Latina y el Caribe [CEPAL], (2020) agrega a esta tipología los datos según su nivel de procesamiento en el curso de la investigación, y distingue:

- Datos primarios o sin procesar: Datos originales que han sido recolectados, pero aún no han sido procesados o analizados. Algunos ejemplos son los registros sonoros, observaciones, notas de campo o datos de experimentos.
- Datos procesados: Datos que han sido digitalizados, traducidos, transcritos, limpiados, validados, verificados y/o anonimizados.
- Datos analizados: Modelos, gráficos, tablas, textos u otros, que han sido creados a partir de los datos primarios y procesados, y que se pretende sean de ayuda en el descubrimiento de información útil, la presentación de conclusiones y la toma de decisiones.

Travieso Rodríguez & Ferreira Araújo (2019) agregan que los datos pueden extraerse en diferentes formatos físicos o electrónicos, e identifican diversas extensiones de archivo, por ejemplo:

- Audiovisuales: .mp3, .wav, .aif, .mp4, .mov, .avi, .mj2, .flac
- Textuales: .txt, .rtf, .bib, .doc, .pdf, .odt, .tex
- Imágenes.jpg, .png, .svg, .tiff
- Gráficos 3D: .obj, .stl, .ply
- Presentaciones: .ppt, .pptx, .pptm
- Geográficos/geoespaciales: .keyhole, .GIS, .gif, .dwg, .shp, .dbf, .mif
- Hojas de cálculo: .csv, .xls, .tsv, .tab, .ods
- Comprimidos: .zip, .rar, .iso
- Estructurados: .mdb, .accdb, .html, .xml

El ciclo de vida de los datos tiene estrecha relación con el ciclo de vida del proyecto de investigación. En cada etapa se pueden generar diferentes tipos de datos o nuevas versiones de los mismos. Es importante entender el ciclo para una adecuada gestión y planificación y su consignación en el denominado Plan de Gestión de Datos (PGD o DMP, por sus siglas en inglés).

La *United States Geological Survey [USGS]*, (2014) propone un modelo lineal y transversal para comprender y gestionar los datos, con el propósito de garantizar que éstos sean detectables, bien descritos y conservados para el acceso y uso más allá de la vida de los proyectos de investigación. El modelo de la USGS consta de seis fases o hitos en la gestión de los datos de investigación:

1. Planificar: tiene por objeto ayudar a los científicos a garantizar la consideración de todas las actividades relacionadas con el manejo de los activos de datos del proyecto, desde el inicio del proyecto, a la publicación y al archivo.
2. Adquirir: representa las actividades a través de las cuales se recogen, generan o consideran y evalúan los datos nuevos o existentes para su reutilización.
3. Procesar: esta fase implica varias actividades relacionadas con la preparación de aportaciones de datos nuevos o previamente recopiladas.
4. Analizar: el análisis representa las actividades relacionadas con la exploración y la interpretación de datos procesados, donde se prueban las hipótesis, los descubrimientos y se sacan conclusiones.
5. Preservar: esta fase recoge el conjunto de actividades relacionadas con el almacenamiento de datos para su uso y acceso a largo plazo.
6. Publicar/Compartir: combina los conceptos de publicación tradicional revisada por pares, con la distribución de datos a través de sitios web, catálogos de datos y medios sociales, entre otros.

Durante la realización de las seis fases de este modelo, paralelamente se producen tres actividades transversales (Fig. 2) que aportan eficacia a los datos:

1. Descripción: incluye la documentación y los metadatos mediante el registro de esta información en cada etapa del ciclo de vida de los datos, en lugar de en las etapas finales. La descripción a lo largo del ciclo de vida ayuda a asegurar la exactitud y comprensión de los datos científicos creados, compilados, procesados y compartidos.
2. Gestión de la calidad: planificar las medidas que garanticen la calidad de los datos al inicio del proyecto y se continúe la vigilancia de esta calidad, durante todo el ciclo.
3. Respaldo y seguridad: implica la gestión de los riesgos físicos de los datos en todo el ciclo de vida de los datos, asegurando al mismo tiempo que los datos sean accesibles.

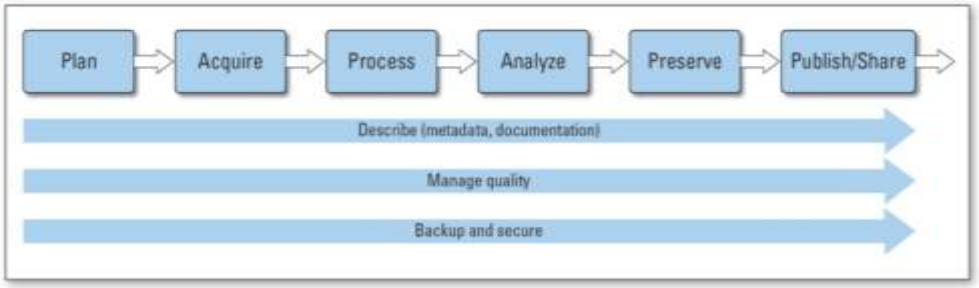

Figura 2: Modelo de ciclo de vida de datos científicos de USGS. Fuente: Faundeen et al., 2014

La *Data Observation Network for Earth – DataONE* (2012) a diferencia de la USGS propone un modelo circular y cíclico que se basa en compartir y reutilizar los datos. El ciclo tiene ocho componentes (Fig. 3):

1. Planificar: descripción de los datos que se compilarán y cómo se gestionarán y serán accesibles a lo largo de su vida.
2. Recoger: las observaciones se hacen a mano o con sensores u otros instrumentos y los datos se colocan en forma digital.
3. Asegurar: la calidad de los datos se asegura mediante comprobaciones e inspecciones
4. Describir: los datos se describen con precisión y minuciosidad utilizando los estándares de la apropiación y la adaptación.
5. Preservar: los datos se envían a un archivo a largo plazo apropiado (es decir, al centro de datos)
6. Descubrir: se localizan y obtienen datos potencialmente útiles, junto con la información pertinente sobre los datos (metadatos)
7. Integrar: los datos de fuentes dispares se combinan para formar un conjunto homogéneo de datos que pueden ser fácilmente analizados
8. Analizar: se analizan los datos

Figura 3: DataONE: Ciclo de vida de los datos. Fuente: Strasser, 2012

El *Digital Curation Center – DDC* (2020) situado en Irlanda, incluye la curación y conservación de datos dentro del ciclo. El DDC creó un modelo que se pude aplicar para la planificación de actividades y asegurarse que se lleven a cabo todas las etapas de forma secuencial, y que recoge cuatro acciones o etapas en el ciclo de vida de los datos:

1. Descripción e Información sobre la representación: asignar metadatos administrativos, descriptivos, técnicos, estructurales y de preservación, utilizando normas apropiadas, para asegurar una descripción y un control adecuados a largo plazo. Recopilar y asignar la información de representación necesaria para comprender y presentar tanto el material digital como los metadatos asociados.
2. Planificación de la preservación: planificar la preservación a lo largo del ciclo de vida de la conservación del material digital. Esto incluiría planes de gestión y administración de todas las acciones del ciclo de vida de la preservación.
3. Vigilancia de la comunidad y participación: vigilar las actividades comunitarias adecuadas y participar en la elaboración de normas compartidas, herramientas y programas informáticos apropiados.
4. Curación y preservación: conocer y emprender las acciones de gestión y administración planificadas para promover la conservación y la preservación a lo largo del ciclo de vida de la conservación

A estas cuatro etapas se añaden las acciones secuenciales que contienen elementos de conservación: planificar las opciones de almacenamiento, crear metadatos de preservación, evaluar los datos y seleccionar para su conservación y preservación a largo plazo, transferir los datos a un archivo, depósito, centro de datos u otro custodio, tomar medidas para asegurar la preservación y retención a largo plazo de la naturaleza autorizada de los datos, almacenar los datos de manera segura, respetando las normas pertinentes, asegurar que los datos sean accesibles tanto para los usuarios designados como para los reutilizadores y crear nuevos datos a partir del original. Además, existen acciones particulares como desestimar los datos que no hayan sido seleccionados para su conservación y preservación a largo plazo de conformidad con políticas, orientaciones o requisitos legales documentados, devolver los datos que no superen los procedimientos de validación para su posterior evaluación y selección, o migrar los datos a un formato diferente (Fig. 4).

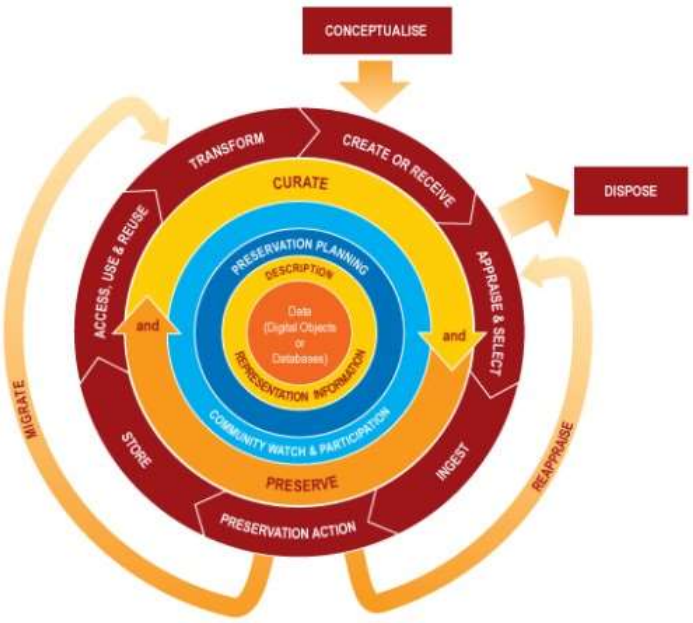

Figura 4: Ciclo de vida de conservación de DCC. Fuente: Higgins, 2008.

2.1.2. Aspectos técnicos de la apertura de datos de investigación

La apertura y la compartición de datos de investigación está vinculada a diversos aspectos técnicos y legales, en este apartado vamos a reflejar algunos de ellos, en particular: los principios FAIR, los planes de gestión de datos, reflexionando también sobre las herramientas al alcance de los investigadores para realizarlos, así como los principales tipos de licencias que permiten abrir y compartir los datos.

Los principios **FAIR** se establecen en el 2016 como un conjunto de directrices para garantizar la reutilización de los datos científicos. Son principios, no estándares que aluden a la "encontrabilidad", que los datos sean localizables *(Findable)*, accesibilidad *(Accessible)*, interoperabilidad *(Interoperable)* y reutilización *(Reusable)* de los datos de investigación. Están diseñados para que los datos, concebidos como objetos digitales puedan ser analizados y procesados por seres humanos y por máquinas. La idea de ser procesable por máquinas se aplica en dos contextos: primero, cuando se hace referencia a los metadatos contextuales que rodean a un objeto digital (¿qué es?), y segundo, al referirse al contenido del propio objeto digital (¿cómo lo proceso / lo integro?). Cualquiera, o ambos, pueden ser procesables por máquina, y cada uno forma su propia forma de analizarlo (Wilkinson et al., 2016). Se trata de aportar información que describa e identifique a los datos y sus instrumentos dentro del proceso de investigación que nos permiten obtenerlos. No solo se centra en los datos en sí mismos, sino en todos los elementos que se usaron para producirlos.

La aplicación de los Principios FAIR no implica necesariamente que los datos deban estar abiertos. La 'A' en FAIR significa "accesible", que esté disponible en condiciones

bien definidas (Mons et al., 2017) bajo una licencia que establezca los términos y alcances de su utilización. Las razones para no permitir el acceso abierto a los datos pueden estar sujetas distintos aspectos de los mismos como son: la confidencialidad y protección de datos sensibles, la seguridad o su uso comercial (patentes). En estos casos en que los datos no estén abiertos o no estén libremente disponibles, sus metadatos deben estarlo, deben ser accesibles para fomentar su reutilización.

Los principios fueron incluidos dentro del Plan Horizonte 2020 como Directrices sobre Gestión de datos FAIR donde se invita a ser parte del *Open Research Data Pilot (ORD pilot)*[6]. Participar en el Piloto ORD no significa necesariamente abrir todos sus datos de investigación. Más bien, el piloto del ODR sigue el principio "tan abierto como sea posible, tan cerrado como sea necesario" y se centra en el fomento de la gestión racional de los datos como parte esencial de las mejores prácticas de investigación (European Commission, 2020). Incluye la elaboración de un Plan de Gestión de Datos (o DMP, Data Management Plan), el ciclo de vida de los datos y la plantilla para la aplicación de los principios FAIR. Durante la primera etapa del piloto de la Comisión, el objetivo fue proporcionar las directrices y manuales para la implementación de los principios en proyectos de Ciencia Abierta en la Unión Europea (UE). En la siguiente etapa, el programa Horizonte Europa 2021-2027 plantea seguir fomentando y exigiendo la gestión FAIR de los datos de investigación, ahora también fomentada la compartición de los datos a través de la mega-infraestructura europea para los datos de investigación conocida como *European Open Science Cloud/Commons (EOSC)*

Los principios FAIR implican una serie de aspectos técnicos (Fig. 5), principalmente metadatos e identificadores persistentes (PIDs) que detallamos, en una explicación simplificada a continuación:

1. ***F**indable* [Localizable], implica asignar un identificador único y persistente (DOI o Handle) e incluirlo en uno de los metadatos enriquecidos y depositarlos en repositorios nacionales e internacionales temáticos o generales.
2. ***A**ccessible* [Accesible], quiere decir que los datos se ejecutan usando un protocolo de comunicación (HTTP(S), FTP o SMTP) de alcance mundial, abierto y gratuito que permite acceder a los metadatos, incluso cuando los datos no estén disponibles.
3. ***I**nteroperable* [Interoperable], supone utilizar vocabularios, formatos, estándares aceptados por la comunidad y vincularlos significativamente con otros metadatos para contextualizar los datos, utilizando identificadores.
4. ***R**eusable* [Reutilizable], tiene que ver con la asignación de licencias claras, sin ambigüedades, que facilite reconocer la procedencia de los datos para que puedan ser citados correctamente. Además de utilizar estándares comunes, aceptados en diferentes disciplinas para enriquecer los datos.

[6] https://data.europa.eu/euodp/data/dataset/open-research-data-the-uptake-of-the-pilot-in-the-first-calls-of-horizon-2020

Figura 5: Principios FAIR. *Fuente: Australian National Data Service (ANDS[7])*

Asimismo, existen múltiples iniciativas, proyectos y comunidades relacionadas con los principios FAIR, para analizar su alcance y facilitar la aplicación de los principios. Las iniciativas más importantes en la actualidad relacionadas con FAIR son GOFAIR y el proyecto europeo FAIRsFAIR, en el que participa la Universidad Carlos III de Madrid.

- GO*FAIR*[8] es una iniciativa que busca poner en funcionamiento los principios FAIR dentro del ecosistema de los datos abiertos, basándose en tres pilares:
  - GO CHANGE: enfocándose en prioridades, políticas e incentivos para implementar *FAIR*. Un cambio sociocultural que involucra a las partes interesadas relevantes en todos los niveles relevantes para el florecimiento de la ciencia abierta.
  - GO TRAIN: coordinando la formación *FAIR* de sensibilización y desarrollo de habilidades. Capacitar a los administradores de datos necesarios capaces de diseñar e implementar planes de gestión de datos adecuados, incluidos datos y servicios FAIR.

[7] https://www.ands.org.au/__data/assets/image/0011/1416098/FAIR-Data-image-map-graphic-v2-721px.png. Créditos de la imagen: The Magnifying glass, Tab, Gears set, Recycle sig, Storage, Infinity, Discussion, Shield, and Man User icons made by Freepik from www.flaticon.com are licensed by CC 3.0 BY. All other icons made by ARDC. Entire FAIR resources graphic is licensed under a Creative Commons Attribution 4.0 International License

[8] https://www.go-fair.org

- GO BUILD: coordinando la tecnología FAIR. Diseñar y desarrollar los estándares técnicos, las mejores prácticas y los componentes de infraestructura necesarios para implementar los principios de datos FAIR.

Los tres pilares son la base para implementar diversas acciones alrededor de GOFAIR. Estas acciones están dirigidas a la "FAIRificación" de los metadatos para que sean entendibles por máquina, los perfiles de implementación, objetos digitales y los financiadores la figura 6 a continuación, resume estas acciones de GOFAIR.

GO FAIR Metadata4Machines	Hackathon que reúnen a expertos que representan a una comunidad de dominio con expertos en metadatos FAIR que los guian sobre las necesidades de datos FAIR de esa comunidad de dominio.
GO FAIR Convergence Matrix	Plataforma que guía sistemáticamente a cualquier comunidad autoidentificada en el proceso de decisión que conduce a implementaciones y prácticas óptimas de FAIR
GO FAIR GOing FAIR DOing FAIR	Es la combinación de los objetos digitales y los datos vinculados. Brinda un estandar minimo definido para la implementación de los principios. El Objeto Digital FAIR garantiza la imparcialidad incluso para los artefactos digitales complejos de componentes múltiples.
GO FAIR Made Easy for Funders	El propósito de este ecosistema emergente de proveedores de servicios FAIR es agilizar la creación de datos FAIR que luego participarán en Internet of FAIR Data and Services (IFDS)

Figura 6: Iniciativas GOFAIR. Fuente: GOFAIR

- FAIRSFAIR[9] es un proyecto que fomenta la práctica de los Principios FAIR de datos en Europa dando soluciones para su utilización desde el inicio de los proyectos de investigación, pasando por el ciclo de vida de los datos. Es un proyecto financiado con fondos europeos, que ha creado un consorcio que cooperará con otros proyectos e iniciativas pertinentes para desarrollar una infraestructura de conocimientos y soluciones prácticas para el uso de los principios de datos FAIR a lo largo del ciclo de vida de los datos de investigación. El proyecto también garantizará una amplia aceptación y una aplicación adecuada de los principios mencionados en la *European Open Science Cloud (EOSC)* por parte de todos los proveedores de datos (CORDIS, 2020).

Los grupos de trabajo FAIRSFAIR (Fig. 7) han desarrollado herramientas y software que ayudan a la evaluación de los profesionales vinculados con la producción y

[9] https://www.fairsfair.eu

administración de datos, así como también los datos en sí mismos y los repositorios que los almacenan.

FAIR Aware	Evaluación en línea con 19 preguntas y ejemplos que permite a los investigadores y administradores de datos conocer su nivel de onocimiento acerca de los Principios FAIR y cómo aplicarlo a sus datos.
FAIR Data Object Assessment Metrics	Evaluación de 17 métricas minimas viabless para la evaluación sistemáticas de los datos FAIR.
F-UJI Automated FAIR Data Assessment Tool	Evaluación en REST basada en 16 de las 17 métricas evaluables de objetos FAIR. Utiliza registros y recursos externos para obtener información sobre las pruebas prácticas implementadas con respecto a las métricas.

Figura 7: Herramientas y software desarrolladas por FAIRSFAIR. Fuente: FAIRSFAIR

Otro aspecto importante en la buena gestión de datos de investigación, desde el punto de vista técnico, pero también de su proyección e implementación, es la creación y seguimiento de un Plan de Gestión de Datos o *Data Management Plan (DMP)* desde el inicio del proyecto que servirá como guía dentro de ciclo de vida de la investigación. El DMP es un elemento clave de la buena gestión de datos y su principal objetivo es describir el ciclo de vida de todos los conjuntos de datos que se van a generar, recoger y procesar durante las actividades realizadas en todos los grupos de trabajo durante el proyecto (Kousouris y Papadopoulos, 2020). Se considera la forma más eficaz de pedirle al investigador que comunique sus intenciones de almacenar, usar, mantener y poner a disposición los datos producidos por su proyecto (Vitale & Moulaison Sandy, 2019).

Son múltiples los beneficios de promover la creación de DMPs entre los investigadores, Smale et al. (2018) los clasifican de la siguiente forma:

- Beneficios profesionales: productividad y visibilidad del investigador.
- Beneficios económicos: aumento del impacto académico y no académico de la investigación, por unidad de inversión en financiación de investigación. Estos beneficios los obtienen el gobierno, la sociedad y los organismos de financiación de la investigación.
- Beneficios institucionales: la utilidad para fines de planificación institucional y cumplimiento.

Cada vez son más las agencias financiadoras que están solicitando presentar un DMP como requisito dentro de la propuesta para obtener financiación para las investigaciones. La estructura del plan estará sujeta al campo y tipo de investigación que se desarrolle, los planes pueden variar y ajustarse a los requerimientos de cada agencia de financiación. Tibo y Jones (2013) afirman que existen dos razones principales para la inclusión de este nuevo requisito: “La primera razón es la

transparencia y franqueza para evidenciar donde se invierten los impuestos de los ciudadanos y comprobar la responsabilidad de los gobiernos en poner énfasis en los verdaderos temas de investigación importantes para el desarrollo del país. La segunda es el retorno de la inversión ya que los datos al ser localizables, accesibles, interoperables y reutilizables genera una mayor rentabilidad de las inversiones y reduce el fraude científico".

Así, la Comisión Europea, en Horizonte 2020, aconsejaba la creación de un DMP que incluyese los principios FAIR, y a partir de este año, Horizonte Europa 2021-2027 este requisito pasa a ser obligatorio para garantizar el acceso abierto a los datos de investigación siguiendo el objetivo propuesto de la Ciencia Abierta: mejor difusión y explotación de los resultados de investigación e innovación, así como apoyo a la participación activa de la sociedad.

Los DMPs deben considerar, al menos, los siguientes aspectos relacionados con los datos de investigación: El manejo de los datos de la investigación durante y después de la finalización del proyecto; Qué datos se recogerán, procesarán y/o generarán; Qué metodología y normas se aplicarán; Si los datos serán compartidos o hechos de libre acceso; Cómo se curarán y conservarán los datos (incluso después de la finalización del proyecto). En el año 2013 el Digital Curation Center presenta, en la versión 4.0, de su lista de verificación para elaborar un DMP. En el primero punto indica que se debe incluir los datos administrativos para identificar y contextualizar los datos. Incluye el código asignado por la agencia o por la institución, el nombre de la agencia de financiación, número, nombre y descripción del proyecto, nombre e identificador persistente de el o los investigadores principales, datos de la persona de contacto, fecha de la primera versión y última actualización del plan y las directrices sobre los datos. Luego se desarrollan las trece preguntas iniciales que deben plantearse al momento de elaborar el plan[10].

Existen múltiples herramientas en línea para crear un DMP siguiendo los requisitos de las principales agencias financiadoras. Una de ellas es DMPTool[11] que fue creada por el Centro de Curación de la Universidad de California (EE.UU.) en el año 2011 para crear planes de gestión de datos en línea y de forma gratuita. Para elaborar un nuevo plan se debe definir qué tipo de proyecto se está elaborando, la organización que respalda la investigación y la agencia de financiación. De acuerdo a la agencia seleccionada se mostrarán las versiones disponibles.

Otra herramienta en línea y gratuita es DMPOnline[12], su interfaz de uso es muy similar a DMPTool. En ambas herramientas se pueden consultar y descargar los planes

[10] http://www.dcc.ac.uk/resources/data-management-plans
[11] https://dmptool.org
[12] https://dmponline.dcc.ac.uk

públicos generados por otros investigadores, las plantillas con los requisitos específicos de las agencias de financiación, el listado de las instituciones participantes, guías de inicio rápido y guías para la gestión de datos para investigadores.

En España, el Consorcio Madroño creo PAGODA[13] un portal con información sobre la gestión de datos de investigación dirigida a investigadores de las Universidades madrileñas. PAGODA incluye PDGOnline que es una adaptación y traducción de DMPOnline del DCC, de acuerdo con los requerimientos del Programa Horizonte 2020 para proyectos que soliciten subvenciones para investigaciones científicas. Se basa en la versión 7 de la plantilla 'PGD Horizonte 2020' proporcionada por *European Commission* (H2020) publicada el 30 de abril de 2020. Al ser una herramienta personalizada el panel de control incluye los Principios FAIR dentro del plan que se estructura en tres tiempos: inicial, detallado y final.

Finalmente, el tercer punto clave en la gestión de los datos de investigación son las **licencias que se le otorgan a los datos**[14]. Es importante definir qué tipo de licencias se utilizarán y cuáles son las condiciones que implica su uso para evitar ambigüedades en la interoperabilidad legal y la reutilización de los datos. Las licencias más usadas son las Creative Commons y las Open Data Commons.

Las licencias Creative Commons[15] son licencias alternativas al copyright que se rigen por cuatro condiciones:

1. Reconocimiento (*Attribution*): Es necesario reconocer la autoría para utilizar la obra.
2. No Comercial (*Non comercial*): No se puede usar la obra para sin fines comerciales.
3. Sin obras derivadas (*No derivate works*): No se puede modificar la obra para crear una obra derivada.
4. Compartir igual (*Share alike*): Se puede crear una obra derivadas siempre que se comparta bajo la misma licencia.

Bajo estas condiciones se pueden elegir hasta seis tipos de licencias diferentes que se resumen en la tabla a continuación (Tabla 2):

[13] https://pgd.consorciomadrono.es

[14] El Portal Europeo de Datos ofrece el servicio de un asistente de licencia dentro de su programa de formación en línea. El usuario puede obtener la licencia de acuerdo a sus necesidades eligiendo el tipo de obligación, permiso y prohibición para su obra: https://www.europeandataportal.eu/es/training/licensing-assistant

[15] https://creativecommons.org

Tabla 2: Licencias Creative Commons. Fuente: CreativeCommons.

Licencia CC	Condiciones
	Permite distribuir, mezclar, adaptar la obra, incluso comercialmente, siempre que le den crédito por la creación original. Es la licencia más abierta.
	Permite remezclar, adaptar y construir sobre la obra incluso con fines comerciales, siempre que le otorguen crédito y licencian sus nuevas creaciones bajo los mismos términos.
	Permite a otros reutilizar el trabajo para cualquier propósito, incluso comercial; sin embargo, no se puede compartir con otros en forma adaptada y se le debe proporcionar crédito.
	Permite a otros remezclar, adaptar y construir sobre su trabajo sin fines comerciales, y aunque sus nuevos trabajos también deben reconocerlo y no ser comerciales, no tienen que licenciar sus trabajos derivados en los mismos términos.
	Permite a otros remezclar, adaptar y construir sobre su trabajo de forma no comercial, siempre que le den crédito y licencian sus nuevas creaciones bajo los mismos términos.
	Sólo permite que otros puedan descargar las obras y compartirlas con otras personas, siempre que se reconozca su autoría, pero no se pueden cambiar de ninguna manera ni se pueden utilizar comercialmente. Es la licencia es la más restrictiva de las seis.

Las licencias Open Data Commons[16] son licencias que permiten aplicar herramientas legales a los datos abiertos. Existen tres tipos de licencias:

1. *Open Data Commons Open Database License (ODbL)*: Permite reproducir, distribuir, usar, adaptar y transformar el conjunto de datos siempre que se reconozca la autoría. Se pueden realizar obras derivadas aplicando la misma licencia de uso. Para versiones con restricciones de uso se debe asegurar el acceso abierto a una copia.
2. *Open Data Commons Attribution License*: Permite reproducir, distribuir, usar, adaptar y transformar el conjunto de datos siempre que se reconozca la autoria y los términos de uso.
3. *Open Data Commons Public Domain Dedication and License (PDDL)*: Permite copiar, distribuir, utilizar, modificar, transformar, producir y construir a partir de la obra.

[16] https://opendatacommons.org

2.1.3. Infraestructuras y datos de investigación

A la hora de resumir las infraestructuras disponibles para depositar datos de investigación y publicarlos en abierto, debemos hablar por un lado de los repositorios de datos de investigación y por otro de EOSC.

Los **repositorios** son una alternativa para dar soporte al depósito los documentos producidos durante una investigación, fueron y son las plataformas elegidas para el desarrollo de la vía verde del Acceso Abierto. Inicialmente se trató de repositorios institucionales con funcionalidades propias de los documentos que albergaban. En general los repositorios de publicaciones o documentos, no están preparados para depositar datos, por eso es habitual que los datos y documentos no estén depositados en el mismo repositorio. Al respecto Hernández-Pérez & García-Moreno (2013) afirman: "El hecho de que sean repositorios independientes no significa que no exista relación entre ellos. Los repositorios de datos de investigación sirven, entre otros fines, para validar resultados de investigación y, por tanto, deben estar vinculados de alguna manera a las publicaciones científicas en donde se muestra para qué fueron utilizados esos datos, por lo que algunos de los problemas se podrán abordar de formar conjunta, tanto para los repositorios institucionales como para los repositorios de datos de investigación" (p.261)

En este sentido, los datos deben ser depositados en repositorios confiables de acuerdo con cada disciplina o en repositorios generales. Se considera confiable si proporciona acceso seguro a largo plazo a los recursos digitales administrados a una comunidad designada, ahora y en el futuro (Research Libraries Group, 2002). OpenAIRE recomienda desde 2017, que el repositorio sea certificado como repositorio digital confiable mediante el sello de aprobación de datos proporcionado por CoreTrustSeal[17] a lo que el proyecto FAIRsFAIR ha añadido además los principios FAIR, creando el concepto de CoreTrustSeal+FAIR (L'Hours et al., 2020)

Existen varias razones para elegir un repositorio u otro como, por ejemplo, en el caso de nuestro estudio, que sea la editorial que publica un artículo, la que recomiende un repositorio particular para depositar los datos. En un estudio elaborado por Alcalá & Anglada (2019) se recogen los siguientes elementos básicos que debería cumplir el repositorio elegido: que proporcione identificadores persistentes, que tenga una alta capacidad de almacenamiento, que soporte diferentes formatos, que cuente con una estrategia de preservación para mitigar riesgos, que garantice la interoperabilidad para el intercambio de datos con otros repositorios de datos, de softwares o con sistemas de gestión de la investigación y que incluyan información sobre la licencia, tipo de citación, esquema de metadatos, versiones, tipos de accesos y difusión en redes sociales académicas (p.14). Algunas opciones de repositorios de datos de la editorial Springer Nature (Scientific Data, 2020) se recogen en la tabla a continuación (tabla 3):

[17] https://www.coretrustseal.org

Tabla 3: Tipos de repositorios. Fuente: SpringerNature

Repositorios Multidisciplinares		
Zenodo	https://zenodo.org	Repositorio creado por OpenAIRE y el CERN con el apoyo de la Comisión Europea
Figshare	https://figshare.com	Repositorio de datos en diversos formatos de archivo
re3data	https://www.re3data.org	Registro global de repositorios de datos de investigación de diferentes disciplinas académicas
Dryad	https://datadryad.org	Repositorio para la publicación de datos y la preservación digital
Biología		
GenBank	https://www.ncbi.nlm.nih.gov/genbank	Colección anotada de todas las secuencias de ADN disponibles públicamente de los National Institutes of Health (NIH)
European Nucleotide Archive (ENA)	https://www.ebi.ac.uk/ena/browser/home	Registro completo de la información de secuenciación de nucleótidos del mundo
DNA DataBank of Japan (DDBJ)	https://www.ddbj.nig.ac.jp/index-e.html	Proporciona servicios de intercambio y análisis de datos de investigaciones de ciencias de la vida y avances en la ciencia
Ciencias Sociales		
Harvard Dataverse	https://dataverse.harvard.edu	Repositorio para compartir, archivar, citar, acceder y explorar datos de investigación
UK Data Service ReShare	https://reshare.ukdataservice.ac.uk	Repositorio del Reino Unido para archivar, publicar y compartir datos de investigación
Open Science Framework (OSF)	https://osf.io	Repositorio abierto y gratuito para respaldar su investigación y permitir la colaboración.
Ciencias Naturales		
Pangea	https://www.pangaea.de	Repositorio abierto para el archivo, publicación y reutilización de datos.
World Data Center for Climate at DRKZ (WDCC)	https://www.dkrz.de	Repositorio de archivo a largo plazo, catalogación, conservación y publicación de datos sobre el clima
Ciencias de la Salud		
ClinicalTrials.gov	https://clinicaltrials.gov	Colección de estudios clínicos financiados con fondos públicos y privados realizados en todo el mundo.

National Database for Clinical Trials related to Mental Illness (NDCT)	https://nda.nih.gov	Archivo para compartir datos de investigación, herramientas, métodos y análisis que permiten la ciencia y el descubrimiento colaborativos
SICAS Medical Image Repository	https://www.smir.ch	Repositorio para almacenar datos de investigación médica

Por otro lado, la EOSC[18] o ***European Open Science Cloud o Commons*** (que es el significado más reciente que se quiere dar a la C)**,** es una infraestructura surgida a iniciativa de la Comisión Europea en el 2015 como una propuesta para crear una infraestructura federada para los datos de investigación europeos bajo los Principios FAIR. Durante la fase inicial se convocó a través de Horizonte2020 la presentación de proyectos que desarrollen modelos para EOSC, esta fase culmino en el 2020. Para la nueva etapa de Horizonte Europa 2021-2027 se pretende posicionar a la UE como líder mundial en la gestión de datos de investigación y garantizar que los investigadores europeos puedan utilizar estos recursos para el desarrollo de la ciencia basada en datos.

La EOSC se crea para proporcionar una plataforma en la nube donde se puedan depositar y procesar datos de investigación financiados con fondos públicos para apoyar la ciencia en la UE. Según la Estrategia Europea de Datos (2020) las etapas para su desarrollo a partir de ahora, son:

- Después de 2020: establecer una estructura de gobernanza de la EOSC renovada, impulsada por las partes interesadas, posiblemente en relación con el lanzamiento de la Asociación Europea de la EOSC correspondiente en el primer trimestre de 2020
- A partir de 2024: abrir, conectar y articular EOSC más allá de las comunidades de investigación, con el sector público en general y el sector privado.
- Para 2025: desplegar operaciones EOSC para servir a los investigadores de la UE.

2.2. Marco regulador para la gestión y compartición de datos de investigación

2.2.1. Directiva de la Unión Europea 2019/1024, de 20 de junio de 2019

La directiva 2019/1024 del Parlamento Europeo y el Consejo de la UE del 20 de junio de 2019 establece el marco normativo para los datos abiertos y la reutilización de la

[18] https://eosc-portal.eu

información del sector público con la finalidad de brindar a la sociedad nuevas formas de conocimiento que permitan su progreso.

Dentro de la regulación se incluyen los conceptos de: datos dinámicos, conjunto de datos de alto valor y datos de investigación. Los datos dinámicos son volátiles y de rápida obsolescencia y pueden actualizarse en tiempo real, generalmente se utilizan en proyectos de Inteligencia Artificial mediante sensores que permiten la descarga de los datos de manera masiva. Por otro lado, los conjuntos de datos de alto valor están delimitados por seis categorías temáticas que se especifican en el anexo I: geoespacial, observación de la tierra y medio ambiente, meteorología, estadística, sociedades y propiedad de sociedades y movilidad. Son considerados de alto valor porque pueden generar servicios de valor añadido, aplicaciones y puestos de trabajo.

Los datos de investigación son identificados como recopilaciones dentro de un proceso de investigación que son aceptados por la comunidad investigadora para validar conclusiones de una pregunta de investigación. Dentro de la tipología se reconocen las estadísticas, resultados de experimentos, mediciones, observaciones derivadas del trabajo de campo, resultados de encuestas, grabaciones e imágenes de entrevistas.

Se recomienda a los estados miembros apoyar acciones y políticas nacionales de acceso abierto que incentiven la disponibilidad de datos de investigación subvencionados por fondos públicos; respetando la propiedad intelectual e industrial, la protección de datos personales, la confidencialidad, la seguridad y los intereses comerciales legítimos. Aplicar los Principios FAIR y establecer un marco general con las condiciones para la reutilización de los datos para lograr la armonización entre las normas y las prácticas nacionales evitando el vacío de información que perjudique el desarrollo de la sociedad de la información.

Además, la directiva hace énfasis en la importancia de la interoperabilidad a través de la pertinente aplicación de los metadatos y el uso adecuado de formatos de publicación legibles y abiertos con el objetivo de lograr la conectividad entre portales de la Unión y la agregación de nuevos metadatos a escala para facilitar la búsqueda lingüística en varios idiomas.

2.2.2. Programa Horizonte 2020 (2014-2020)

El Programa marco Horizonte 2020[19] fue creado por la UE para impulsar el desarrollo de la investigación e innovación durante el periodo 2014 al 2020. Los objetivos estratégicos fueron posicionar a la UE dentro del ámbito científico mundial creando una ciencia de excelencia, proponiendo aumentar la competitividad mediante la implementación de tecnologías y desarrollo de aplicaciones e impulsar las actividades de las pequeñas y medianas empresas centrando su atención en las áreas esenciales para mejorar la vida de los ciudadanos europeos.

[19] https://eshorizonte2020.es

En el punto 6.7 se establece que la difusión de los resultados de los proyectos será a través de publicaciones científicas en acceso abierto revisada por pares y los datos de investigación serán publicados en la iniciativa piloto sobre acceso abierto a datos de investigación (*Open Research Data Pilot*). En ambos casos se deben respetar los derechos de propiedad intelectual, confidencialidad e intereses de los propietarios de los resultados. Además, todas las patentes y publicaciones deben realizar una declaración donde se indique que la investigación fue financiada por la UE y bajo qué acuerdos.

2.2.3. Programa Horizonte Europa 2021-2027

En abril del 2019 se aprobó el próximo programa de inversión en investigación e innovación de la UE: Horizonte Europa 2021-2027 donde se propone reforzar el carácter abierto de los resultados mediante una política de ciencia abierta. Se considera un factor clave dentro de la repercusión científica el fomento de la difusión del conocimiento y la ciencia abierta para el seguimiento del progreso del programa y la cooperación entre los países miembros. Su estructura está compuesta por tres pilares: Ciencia excelente, desafíos mundiales y competitividad industrial europea y Europa innovadora. El primer pilar tiene como objetivo según Gómez (2020) "promover la excelencia científica, crear y difundir conocimiento para atraer talento investigador al espacio europeo y convertir a Europa en un referente mundial en investigación e innovación".

En el artículo 10: ciencia abierta y el artículo 35: explotación y difusión, se explica el planteamiento y las modalidades para lograrlo. Se propone solicitar de forma obligatoria un Plan de Gestión de Datos bajos los principios FAIR usando la premisa: *tan abierto como sea posible y tan cerrado como sea necesario*, tomando en cuenta el apoyo a las habilidades de los investigadores en materia de ciencia abierta, los sistemas de recompensa y el uso de la Nube Europea de la Ciencia Abierta (European Open Science Cloud)[20] como el repositorio de conocimiento para lograrlo.

2.2.4. Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación

La ley 14/2011, de 1 de junio, conocida como "ley de la ciencia" tiene como antecedentes la Ley 13/1986. Fue actualizada veinticinco años después y con un contexto diferente. Para el año 2011 se había logrado la consolidación de una comunidad científica basada en el conocimiento y la innovación, los recursos públicos aumentaron cuantitativa y cualitativamente dando paso a un mayor compromiso con el desarrollo del conocimiento con carácter transversal.

Mediante esta ley, el Sistema Español de Ciencia y Tecnología e Innovación incluye en el artículo 37 la difusión del acceso abierto bajo dos estrategias: La primera es incentivar a las instituciones que forman parte de él a crear repositorios de acceso

[20] https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/eosc_en

abierto que permitan el depósito de las publicaciones de su personal investigador y a su vez propicie la conexión con otras iniciativas similares a nivel nacional e internacional. La segunda es impulsar la publicación obligatoria de la versión electrónica de los artículos, sobre todo de aquellos financiados con fondos de los Presupuestos Generales del Estado, en repositorios del mismo campo de investigación o en repositorios institucionales de acceso abierto. Siempre que el artículo haya sido aceptado en publicaciones de investigación seriadas, en un plazo no mayor a doce meses de la fecha oficial de publicación. Para llevar a cabo ambas estrategias el Ministerio de Ciencia e Innovación posibilita el acceso centralizado a los repositorios y servirá de nexo entre iniciativas afines.

Luego de nueve años de vigencia, el Ministerio de Ciencia e Innovación abrió la consulta pública[21] para la reforma de la ley. Desde el 24 de noviembre hasta el 09 de diciembre del 2020 se podía acceder a la información sobre los cambios significativos que se plantean. El anteproyecto propone dos ejes. El primero es la creación del *ternure track* como una reforma para la contratación fija del personal investigador para su mayor estabilidad laboral. Este tipo de contrato será evaluable cada cuatro años bajo los principios de igualdad, mérito y capacidad. El segundo hace referencia al ámbito de la transferencia del conocimiento y de los resultados de la actividad investigadora como producto del trabajo del personal investigador. En el punto E de este eje se complementa lo expuesto en el artículo 37 de la aún vigente "ley de la ciencia", ya que se continúa incentivando la ejecución de estructuras eficientes para facilitar y fomentar la actividad de transferencia, pero esta vez, se incluyen las entidades vinculadas, incluidas las sociedades mercantiles públicas, como aliadas para lograr este objetivo.

2.2.5. La Estrategia Española de Ciencia, Tecnología e Innovación 2021-2027

La Estrategia Española de Ciencia, Tecnología e Innovación se crea como instrumento de engranaje a las políticas de I+D+I de la UE, aprovechando la sinergia entre ambos programas para lograr la consolidación del Sistema de Ciencia, Tecnología e Innovación de España.

La publicación del último documento para la estrategia de ciencia e innovación en los próximos seis años, coincidió con la pandemia, y por esta razón, incluye actividades dirigidas a resolver los problemas producidos por ella. Partiendo de esta premisa, la estrategia tendrá dos fases: la primera (2021-2023) enfocada a reforzar la carrera investigadora para lograr un recambio generacional óptimo, apoyando las investigaciones en el área médicas y la digitalización. La segunda (2024-2027) consolidar los proyectos de I+D+I y el conocimiento como base para el desarrollo económico y la expansión de recursos (Ministerio de Ciencia e Innovación, 2020).

[21] https://www.ciencia.gob.es/portal/site/MICINN/menuitem.8ce192e94ba842bea3bc811001432ea0/?vgnextoid=0319fc086faf5710VgnVCM1000001d04140aRCRD&vgnextrefresh=1

En el análisis DAFO (Debilidades, Amenazas, Fortalezas, Oportunidades) que recoge el documento, se considera una fortaleza a la creación de repositorios institucionales en centros de investigación y universidades con madurez tecnológica para el desarrollo de Ciencia Abierta, y como una oportunidad al movimiento hacia la Ciencia Abierta, que permitirá mejorar la calidad, la transparencia, el impacto, la reproducibilidad y el acceso de la ciudadanía al conocimiento.

Dentro del modelo de actuación para la Ciencia Abierta, en el eje de actuación 14 se plantea fomentar la divulgación y comunicación científica hacia la sociedad mediante el acceso abierto a los resultados de investigación bajo los principios FAIR en repositorios abiertos. Además, se menciona la inclusión de recursos humanos con formación en la gestión de resultados de investigación para la adecuación de los repositorios y la contribución de datos españoles en la Nube Europea de Ciencia Abierta *(European Open Science Cloud, EOSC).*

2.3. COVID-19 y el cambio de paradigma en la comunicación científica

La pandemia ha ocasionado que las grandes editoriales liberen en acceso abierto sus publicaciones relacionadas con la COVID-19 durante el tiempo que dure el brote del virus. El término Ciencia Abierta es utilizado para definir este movimiento que permite estos cambios en este momento, pero la ciencia abierta no se trata de sólo un momento que se justifica por un acontecimiento mundial (Méndez, 2020), es mantener los avances científicos permanentemente en abierto, tanto las investigaciones como los datos. Las interrogantes que surge ahora son: ¿qué sucederá cuando se encuentre la cura o la vacuna? ¿se cerrará el acceso a estas publicaciones? El intercambio de datos tiene un papel fundamental en las ciencias, aplicándolo al contexto actual esto se evidencia en "la epidemia de ébola de 2014, donde se encontró que el acceso limitado a la información relacionada con el brote podría haber afectado negativamente al final de la crisis. Sin embargo, se aprendió la lección y, en 2016, el problema del virus del Zika se abordó mucho más rápidamente, en parte debido a un mayor uso del intercambio de datos, lo que significó una acción global mucho más coordinada" (Sixto-Costoya et al., 2020, pp.2).

La Unesco (2019) define la ciencia como una de las vías más importantes de acceso al conocimiento. Tiene un papel fundamental del cual se beneficia el conjunto de la sociedad: genera nuevos conocimientos, mejora la educación y aumenta nuestra calidad de vida. Si aplicamos este concepto, la pandemia ha evidenciado que mientras se sigan aplicando las restricciones de acceso a la información, no se está respondiendo eficazmente a la razón de ser de la ciencia. "En estos momentos el mundo está intercambiando información en la búsqueda de una solución... pensando en un descubrimiento para toda la humanidad...recordemos que los avances científicos son bienes comunes" (Amaro, 2020)

Durante la primera ola del virus, se identificaron los cambios iniciales dentro de la comunicación científica afectando a los repositorios temáticos, revistas científicas y repositorios de datos. Los repositorios temáticos se encuentran clasificados dentro de la vía verde del Acceso Abierto (*Green Open Access*) como una opción para el depósito de pos-print y preprint de manera inmediata. Los preprints no pasan por la revisión por pares como en el caso de las revistas científicas, son sometidos a verificación para evitar posibles plagios. Solo los repositorios *bioRxiv*[22] y *medRxiv*[23] revisan los preprint en dos etapas. La primera para verificar que el documento este completo y sin plagios, y en la segunda etapa son examinados por voluntarios académicos o profesionales de la salud. Los tiempos de respuesta para la publicación pueden tomar de dos a cinco días en contextos normales. Los repositorios temáticos más importantes como *bioRxiv, medRxiv, ChemRxiv, arXiv, SocArXiv* y *AfricArXiv* han incluido enlaces directos desde la interfaz principal a la sección especializada en COVID-19 SARS-CoV-2.

Durante los primeros cinco meses del 2020, el número de depósitos de preprints sobre el virus creció de manera desbordante, siendo *medRxiv* y *arXiv* dos de los repositorios con más publicaciones, 2355 y 801 respectivamente.

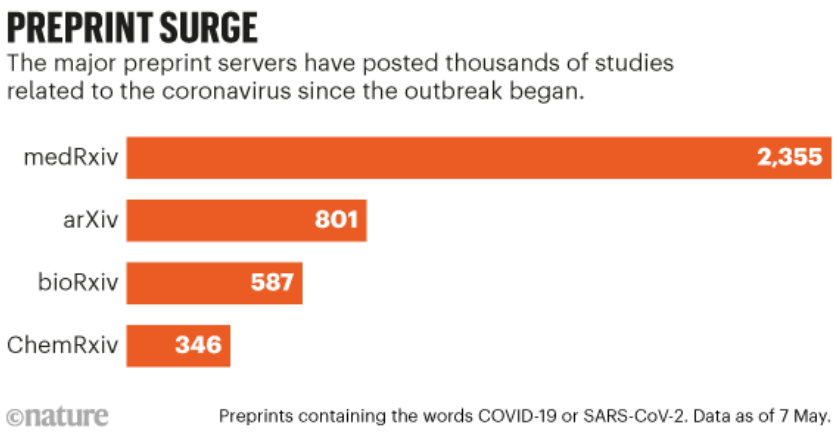

Figura 8: Preprints publicados sobre COVID-19 o SARS-CoV-2 en repositorios temáticos. Fuente: Kwon, D. (2020)

En medio de la pandemia, los repositorios temáticos representan una vía de publicación y diseminación inmediata, porque el depósito de un preprints toma menos tiempo que la publicación de un artículo en una revista arbitrada, pero esta característica puede ser desfavorable dado que no se puede perder la fiabilidad a cambio de la inmediatez y publicar investigaciones que inducen al error, con estudios incompletos o con medias verdades.

[22] https://www.biorxiv.org
[23] https://www.medrxiv.org

Un claro ejemplo de esta situación se aprecia en el preprint "*Uncanny similarity of unique inserts in the 2019- nCoV spike protein to HIV-1 gp120 and Gag*"[24] depositado en el repositorio *bioRxiv* por un grupo de investigadores de la Universidad de Delhi y el Instituto Indio de Tecnología en enero del 2020, en el cual se sugería la creación del virus como parte de una conspiración. El artículo se basaba en las similitudes no fortuitas entre la secuencia genética del SARS-CoV-2 y el VIH-2. La publicación fue tuiteada en la cuenta oficial de Twitter del repositorio y se convirtió en viral en pocas horas. Como consecuencia el preprint fue retirado 72 horas después debido a la alta cantidad de críticas por parte de la comunidad científica.

Analizando este caso, se detectó que se convirtió en viral con más de 20.000 tuits y 56 noticias publicadas en distintos medios de comunicación, pero fue Twitter la red social donde tuvo mayor impacto. En redes sociales académicas el impacto fue bajo. En Google Scholar solo recibió dos citas, una cantidad ínfima a comparación a las 857 citas que recibió el artículo más citado. En Mendeley no alcanzo más de 92 lecturas frente a los casi 2.900 obtenidos por el artículo sobre coronavirus de mayor impacto y en ResearchGate obtuvo 7 recomendaciones casi la mitad que el artículo de mayor repercusión científica sobre el coronavirus (Delgado & Martín-Martín, 2020, p.8).

Para evitar este tipo de casos, *bioRxiv* ya no acepta manuscritos sobre predicciones para el tratamiento del virus basado solo en modelos computacionales. Si un autor con este tipo de manuscritos realiza el depósito, será rechazado y recibirá la recomendación para que lo envíe a una revista para una revisión por pares. Además, luego del incidente con el preprint, colocaron un mensaje en una franja amarilla en todas sus publicaciones donde se puede leer: "bioRxiv está recibiendo muchos artículos nuevos sobre coronavirus SARS-CoV-2. Un recordatorio: estos son informes preliminares que no han sido revisados por pares. No deben considerarse como concluyentes, guiar la práctica clínica / comportamiento relacionado con la salud, ni ser reportados en los medios de comunicación como información establecida".

Otro caso fue el de los investigadores de la Universidad de Nueva York que depositaron en *medRxiv* el preprint "*Factors associated with hospitalization and critical illness among 4,103 patients with COVID-19 disease in New York City*"[25] donde se planteaba el vínculo que existe entre la gravedad del COVID-19 y las personas que sufren de obesidad. La autora principal afirma que los estudios son preliminares y los datos no están completos. A pesar de esto, *The New York Times* (2020)[26] publicó en marzo un artículo basado en el preprint, donde afirmaba que los estudios mostraban que los adultos jóvenes con obesidad tienen más probabilidades de ser hospitalizados por coronavirus, incluso si no tienen otros problemas de salud. El artículo lleva al lector

[24] https://www.biorxiv.org/content/10.1101/2020.01.30.927871v1
[25] https://www.medrxiv.org/content/10.1101/2020.04.08.20057794v1
[26] https://www.nytimes.com/2020/04/16/health/coronavirus-obesity-higher-risk.html

a la confusión porque inicia advirtiendo que los hallazgos son preliminares, que algunos datos aún están incompletos y que el documento no ha sido revisado por pares, pero durante su desarrollo toma como verdadero lo planteado y se apoya en entrevistas realizadas a médicos y especialistas en el tema.

Por otro lado, el profesor Jean-François Delfraissy, que dirige el Consejo Científico que asesora al gobierno francés sobre la epidemia, afirmó en una entrevista que el virus puede afectar a los jóvenes, en particular a los jóvenes obesos, publicado también en un preprint (De Clercq, 2020)

Una iniciativa para mitigar este tipo de casos es *Outbreak Science Rapid PREreview*[27], una plataforma para encontrar, solicitar y revisar preprints relacionados con el brote del virus. Este proyecto es financiado *Wellcome Trust*, organización que elaboro la declaración, con la colaboración de dos organizaciones sin fines de lucro como son *Outbreak Science*[28] y *PREreview*[29].

El proceso de publicación de las revistas científicas también se ha ido adaptando durante la pandemia. Los tiempos de respuesta de la etapa de revisión y de edición se han acortado debido al incremento de las investigaciones. Según el investigador Daniel Torres (2020) estamos frente a una avalancha de estudios científicos, en el 2004 se publicaron una media de 3.000 artículos sobre coronavirus al año. Ahora mismo, se publican 700 cada día. Llevamos unos 20.000 en tres meses.

En el estudio elaborado por Horbach (2020), antes y durante la pandemia, determinó que las revistas académicas lograron disminuir la duración de su proceso de publicación en un 49%, o 57 días en promedio… esta aceleración se refiere tanto a la etapa de revisión (entre la presentación y la aceptación) como a la etapa de edición (entre la aceptación y la publicación). Las revistas estudiadas acortaron ambas etapas en un 47% (45 días) y un 61% (14 días) respectivamente.

La revista *Nature,* y su editor *Springer Nature,* en su editorial del 24 de marzo de 2020 se comprometieron a compartir abiertamente y lo más rápido posible sus investigaciones. Convocaron a todos los profesionales expertos en el tema para unirse como revisores. Este compromiso se reafirma en su editorial del 17 de abril del mismo año, donde se indica cuáles son los cambios dentro de su proceso de arbitraje. Por ejemplo, los envíos iniciales fueron respondidos dentro de las 24 horas en la mayoría de los casos. El proceso de revisión se agilizo gracias a la incorporación de nuevos revisores, lo que hace posible llegar a una primera decisión posterior a la revisión dentro de los 10 días calendario desde la presentación del manuscrito. Si la

[27] https://outbreaksci.prereview.org
[28] https://outbreakscience.org
[29] https://prereview.org

investigación aún está en fase de planificación o recopilación de datos se puede consultar previamente con la revista.

En la investigación realizada por Eisen et al., (2020) se plantearon los principales cambios ocurridos en el proceso de publicación de la revista *eLife*. Se redujeron las solicitudes de experimentos adicionales durante las revisiones como normalmente se venía haciendo y se exhortó a los editores a que acepten sin demora los manuscritos enviados. Se suspendió el límite de dos meses para las revisiones. Los preprints de los artículos deben estar depositados en *bioRxiv* o *medRxiv* e implementar un plan para invitar a investigadores que recién se inician en la carrera investigadora a ser parte del proceso de revisión.

Brock (2020) afirma que la revista *Royal Society Open Science* reestructuró sus tiempos de revisión para sus *Registered Reports (RRs)*. Los informes registrados se basan en aplicar la revisión por pares antes de la recopilación de datos para evitar que los investigadores sigan pistas falsas, dando énfasis a los métodos y plan de análisis que se seguirán en el estudio. Este proceso normalmente tiene un costo, pero para maximizar la accesibilidad a la convocatoria se han eliminado automáticamente los cargos.

Los cambios incluyen, además de eliminar los pagos, acortar los tiempos que toma la revisión inicial de la primera etapa en siete días posteriores a la recepción de la presentación. Es la primera vez que se establecen estos tiempos mínimos para las revisiones; por este motivo la revista inicio la convocatoria para revisores que no necesariamente deben ser virólogo o epidemiólogo porque las investigaciones involucran otros campos del conocimiento como matemáticas, estadística, psicología y ciencias sociales. También se convocó a los autores a enviar sus informes mediante un cuestionario en línea donde se indique la pregunta de investigación, las hipótesis, planes de análisis y la interpretación prospectiva.

Con respecto a los datos abiertos se han convertido en un recurso clave en este contexto, según su tipología contribuye al análisis y desarrollo de diagnósticos específicos, son base para la transparencia de los gobiernos y toma de decisiones para los ciudadanos. Wu et al. (2020) fueron los primeros en publicar la primera secuencia del ADN de la COVID-19. Los investigadores de la Universidad de Fudan en Shangai (China) realizaron la publicación en el repositorio de datos *GenBank*. Incluyeron datos de secuencia, anotaciones y metadatos del virus de un paciente que se cree que contrajo la enfermedad en un mercado de mariscos de la provincia de Hubei (China).

A nivel mundial, los gobiernos de cada país están publicando sus datos abiertos sobre la pandemia. La Comisión Europea creó *COVID-19 Data Portal*[30] para facilitar el intercambio y análisis de datos para acelerar la investigación. En Latinoamérica, los

[30] https://www.covid19dataportal.org

gobiernos de Perú[31], Brasil[32], México[33], Chile[34] y Colombia[35] publicaron sus datos en formatos adecuados para su reutilización. Por su parte, el Center *for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)(EE.UU.)*[36] está publicando los datasets epidemiológicos a nivel mundial.

La característica presente en todos los casos mencionados anteriormente es la inmediatez, ésta se justifica en la necesidad mundial por encontrar un remedio que frene la propagación del virus. Para hallar una solución es necesario que los artículos y sus datos sean verificados previamente por un grupo de científicos que validen los resultados de las investigaciones. En nombre de la inmediatez no debería publicarse investigaciones sesgadas o basadas solo en algoritmos matemáticos que conducen a los investigadores por caminos ciegos y sin salida que dificultan el quehacer científico.

La publicación de los resultados de una investigación es difundida no sólo entre los investigadores, sino a la sociedad en general mediante los medios de comunicación. En algunos casos la información que se transmite es falsa o carente de validez, produciendo desconcierto y pánico entre la población. No solo los ciudadanos son inducidos al error, las autoridades con cargos oficiales han tenido que rectificar sus declaraciones basadas en este tipo de noticias.

2.4. Iniciativas Open Science para publicaciones y datos sobre la COVID-19

2.4.1. *Manifesto for European Union COVID-19 Research*

La Comisión Europea publicó en julio del 2020 el manifiesto para maximizar la accesibilidad a los resultados de las investigaciones sobre la lucha contra la COVID-19[37] que busca el apoyo y adhesión voluntaria de los agentes públicos y privados que obtuvieron financiación de la UE, así como de instituciones y personas dedicadas a la investigación.

El manifiesto propone publicar las investigaciones, resultados y datos que fueron subvencionados en la Plataforma de Resultados Horizon y la Plataforma de Datos Europeos sobre Covid-19[38] para hacerlos públicos y accesibles lo más pronto posible siguiendo los principios FAIR. Siendo los más importante asignar licencias no exclusivas y libres de regalías bajo el compromiso de los licenciatarios a distribuir bajo las mismas

[31] https://www.datosabiertos.gob.pe/search/field_topic/covid-19-917
[32] https://covid.saude.gov.br
[33] https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico
[34] https://www.gob.cl/coronavirus/cifrasoficiales/#resumen
[35] https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data
[36] https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6
[37] Maximising the accessibility of COVID-19 research results: https://ec.europa.eu/info/research-and-innovation/research-area/health-research-and-innovation/coronavirus-research-and-innovation/covid-research-manifesto_en
[38] https://www.covid19dataportal.org

condiciones los productos o servicios que ayuden a prevenir, diagnosticar, tratar o contener el virus. Las licencias estarán abiertas hasta un año después que la Organización Mundial de la Salud declarase la COVID-19 como una emergencia de salud pública de interés internacional con un plazo máximo del 01 de enero del 2022.

2.4.2. Compartir datos de investigación y hallazgos relevantes para el nuevo brote de coronavirus (NCOV)

El intercambio de información en busca de la cura y el control sobre la pandemia hace necesaria la difusión de investigaciones confiables y datos sólidos en el menor tiempo posible. Bajo esta premisa, la organización benéfica de investigación *Wellcome Trust* (2020) hizo un llamamiento a investigadores, revistas y financiadores para garantizar que los resultados de las investigaciones, así como los datos relevantes sobre el brote del virus se compartan rápida y abiertamente para informarlo a la salud pública y garantizar que la Organización Mundial de la Salud (OMS) tenga acceso rápido a los hallazgos emergentes, que podrían ayudar a salvar vidas en el mundo.

La declaración fue firmada por más de setenta instituciones entre grandes editoriales como *Nature, Cell, The Lancet, Taylor&Francis, Willey, Elsevier, Emerald Publishing* e instituciones internacionales de salud, investigación y educación. Todos los firmantes se comprometen a publicar todas sus investigaciones revisadas por pares relevantes en acceso abierto o disponible gratuitamente al menos durante la duración del brote. Los resultados de las investigaciones se comparten de inmediato con la OMS después de la publicación de la revista, esto lo realizará la revista con consentimiento del autor. La publicación de preprints se realizará en repositorios de acceso abierto antes de la revisión por pares, siempre que cuenten con políticas claras sobre la disponibilidad de los datos depositados. Los datos de investigación provisional y final; así como los protocolos y estándares utilizados para recopilarlos serán compartidos de la forma más rápida y amplia posible incluyendo las comunidades científicas, la OMS y la salud pública. Los autores tienen claro que los datos o preprints compartidos antes del envío no impide su publicación en estas revistas.

2.4.3. *Data Together COVID-19*

En marzo 2020, las principales organizaciones internacionales que trabajan sobre los datos de investigación publicaron una declaración donde se comprometen a trabajar de forma conjunta para optimizar el ecosistema de los datos de investigación mundial en medio de la pandemia de la COVID-19. CODATA[39], GO FAIR[40], RDA[41] y WDS[42] son las cuatro organizaciones que se unen en la iniciativa *Data Together COVID-19 appeal and actions*[43], para implementar un ecosistema FAIR que brinde una infraestructura capaz

[39] CODATA: Committee on DATA (International Science Council): https://codata.org
[40] GOf FAIR: https://www.go-fair.org
[41] RDA: Research Data Alliance: https://www.rd-alliance.org
[42] WDS: World Data System: https://www.worlddatasystem.org
[43] Más información sobre esta iniciativa: https://codata.org/data-together-covid-19-appeal-and-actions

de visualizar, unir y reutilizar datos provenientes de diferentes fuentes sin dejar de lado la protección de datos y permitir la descentralización y análisis asistido por máquina.

En el contexto de la pandemia generada por el virus COVID-19 las plataformas construidas en base a la ciencia abierta y los principios FAIR deben tomar en cuenta las siguientes consideraciones:

- El acceso a los datos sensibles y personales debe ser restringido de diversas maneras, y las instituciones que gestionan estos conjuntos de datos no pueden liberarlos abiertamente sin procesamiento adicional o controles de acceso.
- El almacenamiento de los datos no puede ser el único enfoque que se aplique a los datos, debe complementarse con la generación de metadatos enriquecidos y legibles por máquinas para potenciar su análisis o visualización.
- La cantidad de datos generados en este contexto es alta y genera la difusión de información falsa que se da por verdadera. Debería existir un mecanismo para mitigar esos peligros.
- Facilitar y mejorar los métodos de aprendizaje distribuidos para fijar que los algoritmos y los servicios que se ofrecen puedan interactuar con los metadatos y datos FAIR.
- Es necesaria una amplia comunidad de expertos de confianza para revisar las reclamaciones nuevas y existentes relevantes para las intervenciones de COVID-19

El ecosistema debería ser monopolizado por ninguna organización pública o privada, para FAIR no evitarlo debe existir un control de calidad y un plan de certificación mínimo para todos los componentes.

3. Análisis de los datos de investigación de las publicaciones españolas sobre COVID-19: Resultados y Discusión

Tal y como hemos definido en el apartado dedicado a la metodología, para analizar el estado de publicación de datos de investigación vinculados a artículos científicos españoles sobre la COVID-19 analizamos los 8 primeros meses de la pandemia, seleccionando una muestra final de 306 artículos, donde 219 (72%) no contaba con datos de investigación asociados. El 28% restante tiene los datos de investigación vinculados, de alguna forma a la publicación: 15 (5%) artículos publicaron sus datos de investigación en repositorios, 20 (6%) no tiene los permisos para la publicación de los datos y 52 (17%) presentan los datos (u otros resultados) como “material complementario”.

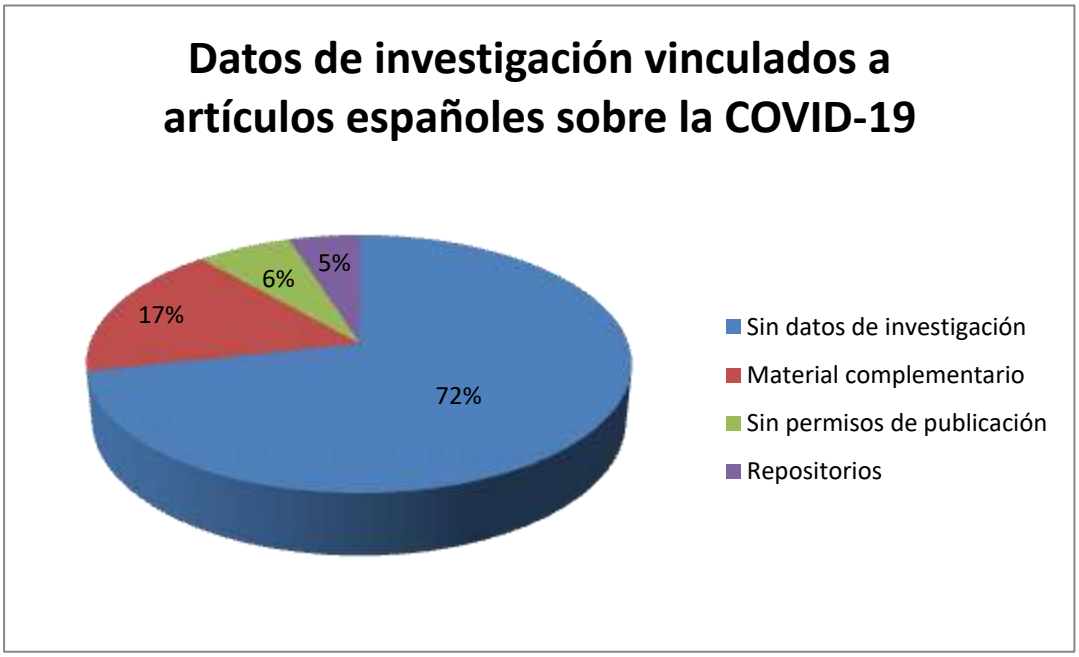

Figura 9: Datos de investigación vinculados a artículos españoles sobre la COVID-19. Fuente: Elaboración propia.

Centramos nuestro análisis en ese 28% de la muestra que se divide en los tres grupos de artículos que hemos señalado: artículos con datos depositados en repositorios, artículos con datos asociados, pero sin permisos de publicación y, artículos que tienen material complementario.

El primer grupo de artículos con datos asociados fue analizado aplicando los nueve atributos que construimos al elaborar la metodología de estudio[44]: 1. disponibilidad, 2. accesibilidad, 3. formato, 4. licencia, 5. vinculación, 6. financiación, 7. política editorial, 8. contenido y 9. estadísticas. Para el primer atributo de disponibilidad se identificó cuál era el repositorio donde se habían depositado los datos (Fig. 10). Los repositorios que se utilizaron para la publicación y disponibilidad de los datos de los 15 artículos identificados en este grupo, fueron: ClinicalTrials (4), Figshare (4), GitHub (2), Mendeley Data (2) GEO -Gene Expression Omnibus (1), Zenodo (1) y el repositorio institucional de la universidad de afiliación de los autores (Universidad de Barcelona, UB) (1).

En general, la elección del repositorio debería depender o depende de la disciplina científica donde se desarrolle la investigación. Si no se cuenta con un repositorio temático puede utilizarse uno general como Zenodo, Mendeley o Figshare. Hemos trabajado y experimentado con todas las plataformas resultantes en este estudio y cada una de ellas tiene un flujo de trabajo diferente para el depósito y publicación de los datos. En el caso de ClinicalTrials se deposita los datos mediante un formulario que

[44] Ver Tabla 1, p. 11.

genera los datos al momento de exportarlos de acuerdo a los formatos que permite el repositorio. En el caso de GitHub, se depositan los algoritmos y códigos utilizados, se toma especial cuidado en la estructura de los archivos que van acompañados de algunos casos de un archivo en formato .txt "léeme" para comprender el contexto de los datos, pero no se tiene mayor precisión en los metadatos que deberían describirlos. Para Zenodo, Figshare y Mendeley se podría decir que es el caso contrario, se introducen los metadatos que describan al conjunto de datos, pero los archivos no tienen una estructura que permita entenderlos o un archivo de lectura que contextualice la investigación. No hay que olvidar que el investigador, es quien está en medio de estas alternativas y que debe invertir tiempo en capacitarse en el manejo de cada una, tiempo que muchas veces no tiene planificado y que se convierte (como hemos podido comprobar en algunos casos) en conjuntos de datos incompletos, mal depositados o no depositados.

Figura 10: Datos depositados en repositorios especializados y generales. Fuente: Elaboración propia

Los atributos dos y cinco (accesibilidad y vinculación) fueron analizados en conjunto, ya que son características de los datos estrechamente relacionadas, ya que para vincular un artículo con sus datos es necesario que ambos recursos dispongan de un identificador único y persistente. Todos los artículos (15) tienen asignados un identificador persistente de objeto digital, DOI (Tabla 4). En el caso de los datos: en 7 casos de los 15, los conjuntos de datos tienen asignado un DOI que identifica el dataset, 1 posee un handle, y los 7 restantes disponen de un enlace de acceso que no es un identificador persistente (PID) que permita asegurar la accesibilidad y perdurabilidad. La vinculación entre documento/artículo y sus datos subyacentes, se

cumple sólo en nueve de los casos, mientras que en seis no hay ninguna vinculación entre datos y artículo.

Es importante destacar que la existencia de un identificador persistente (PID) es fundamental para que los datos sean considerados FAIR, concretamente en relación a la localización y la accesibilidad, ya que contar con un DOI, posibilita el acceso permanente y propicia también la localización de los datasets de manera eficaz. Además, el PID es fundamental para poder citar los datos y para su reutilización. Según comentamos en el estudio del estado del arte, una de las características básicas a la hora de elegir un repositorio para los datos de investigación es que éste soporte la asignación de DOIs para los datos.

Para los datos que fueron publicados en un repositorio específico, éstos tienen un DOI propio para los datos, pero en el caso de los documentos donde los datos se asocian como "material complementario", los datos tienen el mismo identificador que el documento, con lo cual no se podría ni localizar ni citar separadamente. La cita según Assante et al. (2016) es "un mecanismo clave en la publicación de datos de investigación, ya que permite a los propietarios de datos obtener el reconocimiento adecuado para publicar sus conjuntos de datos y a los consumidores de datos para referirse explícitamente a los conjuntos de datos que (re) utilizan en su investigación" (p. 17)

Tabla 4: Accesibilidad y vinculación de los artículos y sus datos. Fuente: Elaboración propia.

DOI del artículo	Repositorio de datos	PID[45]	Enlace	Vinculación
10.3389/fpubh.2020.00243	ClinicalTrials	No	https://clinicaltrials.gov/ct2/show/NCT04326790	Si
10.1016/j.jacc.2020.05.040	ClinicalTrials	No	https://clinicaltrials.gov/ct2/show/NCT04321278	Si
10.1038/s41562-020-0908-8	ClinicalTrials	No	https://clinicaltrials.gov/ct2/show/NCT04326790	Si
10.1371/journal.pntd.0008065	ClinicalTrials	No	https://clinicaltrials.gov/ct2/show/NCT03201185	Si
10.1016/j.clim.2020.108486	Figshare	DOI	https://doi.org/10.25384/SAGE.12294539.v1	Si
10.1177/1747493020923234	Figshare	DOI	https://doi.org/10.3389/fpubh.2020.00243.s001 https://doi.org/10.3389/fpubh.2020.00243.s002	Si
10.1038/s41467-020-17318-x	Figshare	DOI	https://doi.org/10.6084/m9.figshare.12850712.v1	Si

[45] PID hace referencia al identificador persistente (Persistent Identifier, PID, por sus siglas en inglés) del *dataset* o conjunto de datos relativo a la publicación analizada.

DOI del artículo	Repositorio de datos	PID[45]	Enlace	Vinculación
10.1016/S2352-3018(20)30164-8	Figshare	DOI	https://doi.org/10.1371/journal.pntd.0008065	Si
10.1001/jamanetworkopen.2020.13136	GEO	No	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147863	Si
10.1186/s12916-020-01619-5	Zenodo	DOI	http://doi.org/10.5281/zenodo.3701751	No
10.1089/big.2020.0051	GitHub	No	https://github.com/DataLabUPO/CVOA_academic https://github.com/DataLabUPO/CVOA_LSTM	No
10.1080/14767058.2020.1793320	GitHub	DOI	https://github.com/CSSEGISandData/COVID-19	No
10.1016/j.hjc.2020.03.002	Mendeley Data	DOI	http://dx.doi.org/10.17632/98xcryjtnc.4	No
10.1016/j.medcli.2020.04.004	Mendeley Data	DOI	http://dx.doi.org/10.17632/3xzm276pvk.1	No
10.1016/j.cell.2020.04.004	Rep. Dig. UB	Handle	http://diposit.ub.edu/dspace/handle/2445/56611	No

El tercer atributo está relacionado con el tipo de formato de los datos. Los datos fueron depositados en diferentes formatos de acuerdo a su tipología (Tabla 5). En los casos de ClinicalTrials y GEO fueron registrados de acuerdo al formulario y formato propio del repositorio. En el caso de ClinicalTrials pueden descargarse en .pdf, .xml, .txt, .csv y para GEO se pueden descargar todos los archivos en una carpeta comprimida .tar en los formatos SOFT, MINiMAL y .txt. En Mendeley Data fueron depositados en .xlsx. Zenodo en .txt. Figshare en .docx, .pdf y en GitHub en .csv, .java y .pyc. Asimismo, los datos que son depositados en formatos cerrados o propietarios como .doc, .pdf, .xls están sujetos a la continuidad del software que los produce y de sus actualizaciones bajo pago o en un futuro llegar a la obsolescencia. Es una buena práctica utilizar formatos abiertos producidos por software gratuitos que permitan su reutilización.

Tabla 5: Tipos de datos abiertos y formatos de archivos. Fuente: Elaboración propia.

Tipo de datos	Formatos
Estructurados	.xml, .java, .pyc
Textuales	.txt, .pdf, .doc, .docx
Hojas de cálculo	.xls, .xlsx, .csv
Comprimidos	.tar
Experimentales	.soft, .minimal

El cuarto atributo para nuestro análisis es la licencia y las condiciones que las rigen. En este sentido, en la tabla 6 se pueden visualizar que Zenodo y Mendeley Data utilizan en todos los registros consultados las licencias Creative Commons Atribución 4.0 Internacional (CC BY 4.0) que permite distribuir, mezclar, adaptar la obra, incluso comercialmente, siempre que le den crédito por la creación original. Para los conjuntos de datos depositados en Figshare, tres datasets tienen licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0), mientras que un datasest tiene asignado directamente Copyright. ClinicalTrials, dentro de sus términos y condiciones afirma que los datos depositados en esta plataforma son de dominio público, pero se solicitan se reconozca la fuente de información, si se hace uso o reproducción de los datos. GEO dentro de los descargos de responsabilidad *(disclaimer),* sostiene que NCBI *(National Center for Biotechnology Information)*, la institución responsable de la plataforma, no impone restricciones al uso o distribución de los datos; a menos que se indique lo contrario, los documentos y archivos en los servidores web se pueden descargar y reproducir libremente. En GitHub, uno de los registros no indica la licencia y el otro utiliza la licencia pública GNU donde se estipula que todo el mundo está autorizado a copiar y distribuir copias textuales de este documento, pero no se permite cambiarlo. El Repositorio Digital de la Universidad de Barcelona aplica la licencia CC0 1.0 Universal (CC0 1.0) que cede la obra al dominio público, mediante la renuncia a todos sus derechos a la obra bajo las leyes de derechos autorales en todo el mundo, incluyendo todos los derechos conexos y afines, en la medida permitida por la ley. La tabla 6, a continuación recoge la dispar casuística de las licencias de los 15 datasets vinculados a artículos españoles sobre COVID-19 estudiados.

Tabla 6: Licencias aplicadas a los datos. Fuente: Elaboración propia.

Repositorio	Enlace	Licencia
Zenodo	http://doi.org/10.5281/zenodo.3701751	Atribución 4.0 Internacional (CC BY 4.0)
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04326790	Dominio público.
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04326790	Dominio público.
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT03201185	Dominio público.
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04321278	Dominio público.
Figshare	https://doi.org/10.1371/journal.pntd.0008065	Atribución 4.0 Internacional (CC BY 4.0)
Figshare	https://doi.org/10.25384/SAGE.12294539.v1	http://rightsstatements.org/vocab/InC/1.0/
Figshare	https://doi.org/10.3389/fpubh.2020.00243.s001 https://doi.org/10.3389/fpubh.2020.00243.s002	Atribución 4.0 Internacional (CC BY 4.0)

Repositorio	Enlace	Licencia
Figshare	https://doi.org/10.6084/m9.figshare.12850712.v1	Atribución 4.0 Internacional (CC BY 4.0)
GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147863	Dominio público
GitHub	https://github.com/CSSEGISandData/COVID-19	GNU Affero General Public License v3.0
GitHub	https://github.com/DataLabUPO/CVOA_academic https://github.com/DataLabUPO/CVOA_LSTM	No indica
Mendeley Data	http://dx.doi.org/10.17632/3xzm276pvk.1	Atribución 4.0 Internacional (CC BY 4.0)
Mendeley Data	http://dx.doi.org/10.17632/98xcryjtnc.4	Atribución 4.0 Internacional (CC BY 4.0)
Repositorio Digital de la UB	http://diposit.ub.edu/dspace/handle/2445/56611	CC0 1.0 Universal (CC0 1.0)

El sexto atributo estudiado, responde a las agencias u organismos financiadores que apoyaron la realización de las investigaciones. En la muestra de artículos sobre COVID19 , se encontraron reconocidas 26 agencias (Tabla 7), de las cuales el Instituto de Salud Carlos III fue el que más proyectos tuvo bajo su subvención (4), seguido de la Generalitat de Catalunya (3), H2020 Marie Skłodowska-Curie Actions (2) y el Ministerio de Ciencia e Innovación (2).

Es importante mencionar que las agencias financiadoras durante la aplicación del Plan Horizonte 2020 no incluían como requisito obligatorio la presentación de un PGD del proyecto, y la publicación de los datos si bien es altamente recomendada, los autores podían elegir la opción de no hacerlo *(opt-out)* por diversas razones, como por ejemplo la privacidad o la explotación comercial de los datos, entre otras. Sin embargo, como analizamos anteriormente en el programa marco Horizonte Europa este requisito será obligatorio para la obtención de un fondo de financiación.

Tabla 7: Agencias financiadoras y número de proyectos. Fuente: Elaboración propia

Agencia Financiadora	Nº de proyectos
Instituto de Salud Carlos III	4
Generalitat de Catalunya	3
European Regional Development Fund	2
H2020 Marie Skłodowska-Curie Actions	2
Ministerio de Ciencia e Innovación	2
Agencia Estatal de Investigación	1
Edwards Lifesciences	1
European Commission	1
European Research Council	1

Agencia Financiadora	Nº de proyectos
Fundació Privada Daniel Bravo Andreu	1
Fundación Alfonso Martín Escudero	1
Fundación Científica Asociación Española Contra el Cáncer	1
Hellenic Society of Cardiology	1
Horizon 2020 Framework Programme	1
Laboratorios Abbott	1
Ministerio de Economía y Competitividad	1
University Grants Commission	1
Vetenskapsrådet	1

El séptimo atributo de estudio tiene relación con las editoriales y sus políticas en relación a la publicación de datos de investigación vinculados a los artículos que publican. En la tabla 8 podemos identificar que las editoriales Elsevier y Nature Research tienen más artículos publicados con datos asociados. De los 15 artículos analizados, ambas editoriales tuvieron 3 y 2 artículos respectivamente.

Tabla 8: número de artículos publicados con datos asociados por editorial. Fuente: Elaboración propia.

Editorial	Nº artículos
Elsevier B.V.	3
Nature Research	2
Academic Press Inc.	1
BioMed Central Ltd.	1
Blackwell Publishing Ltd	1
Cell Press	1
Ediciones Doyma, S.L.	1
Hellenic Cardiological Society	1
NLM (Medline)	1
Public Library of Science	1
SAGE Publications Ltd	1
Springer	1

Respecto a las políticas de datos de investigación, todas tienen en común que no establecen como requisito obligatorio el depósito de los datos (tabla 9). Elsevier alienta a los investigadores a publicar sus datos y les sugiere un alojamiento gratuito en Mendeley Data. Los datos pueden ser enviados durante el proceso de envío del manuscrito o depositados directamente, en ambos casos se les asignará un DOI y una licencia Creative Commons. Springer Nature divide sus políticas en cuatro tipos, que dependen de las características propias de los datos: citación, intercambio, disponibilidad y revisión por pares Cuenta con un servicio de soporte de datos que consiste en la organización de los archivos en una estructura lógica, edición de

metadatos, embargos acceso privado y anónimo Se recomienda hacer el depósito sin costo en Figshare si el peso de los datos es igual o menor a 50 GB. En el caso que sea mayor a 50 GB el costo va desde 390 euros hasta 4,000 euros (1TB). Cell Press recomienda cargar los datos en Mendeley Data, GitHub o Dryad para aumentar la veracidad de la investigación y evitar problemas con los resultados de la investigación por la falta de acceso a los datos originales, que puede llegar a ser motivo de retractación. PLOS y SAGE también animan a los autores a publicar sus datos, pero no sugieren a un repositorio en especial, hacen énfasis en las restricciones éticas y legales que acompañan a los datos y la declaración de accesibilidad que debe incluirse. En el caso de Taylor and Francis mencionan que los datos deben cumplir con los estándares FAIR e ir acompañados de metadatos y descriptores para que el intercambio sea significativo y útil para otros investigadores.

Aunque las editoriales alienten a los investigares a publicar sus datos, esto no significa que los repositorios generales se conviertan en el cajón de sastre de la ciencia, donde se depositen documentos que no responden a la definición de datos de investigación y en cualquier formato sin ninguna restricción al respecto. Es ese sentido, es necesaria la validación o verificación desde los propios editores o repositorios de los conjuntos de datos a los que dan acceso… para asegurar que son efectivamente datos de investigación originales (Travieso Rodríguez & Ferreira Araújo, 2019, p. 14)

Tabla 9: Política de datos de investigación por editorial. Fuente: Elaboración propia.

Editorial	Política de datos de investigación
Elsevier	No es obligatorio compartir los datos de investigación, pero alientan y apoyan a los autores brindándoles herramientas y servicios que permitan hacerlo. Sugieren a Mendeley Data como una opción de almacenamiento y conservación de sus datos aplicando las licencias CreativeCommons y HiveBench para el flujo de trabajo de investigación.
Springer Nature	Se alienta a los autores, cuando sea posible y aplicable, a depositar datos que respalden los hallazgos de su investigación en un repositorio público. Los autores y editores que no tengan un repositorio preferido deben consultar la lista de repositorios y la política de datos de investigación.
Cell Press	Se recomienda se cargue los datos originales y scripts de programación en Mendeley Data, Dryad, GitHub u otros repositorios, ya que esto aumenta la confianza del lector en los hallazgos. Si surgen problemas con los resultados de la investigación, la falta de acceso a los datos originales dificulta hallar una solución y puede ser motivo de retractación.

Editorial	Política de datos de investigación
Public Library of Scienc e - PLOS	Se requiere que los autores pongan a disposición del público todos los datos necesarios para replicar los hallazgos de su estudio sin restricciones en el momento de la publicación. Cuando las restricciones legales o éticas específicas prohíben el intercambio público de un conjunto de datos, los autores deben indicar cómo otros pueden obtener acceso a los datos.
SAGE Publications Inc.	Se anima a los autores a compartir sus datos de investigación y, cuando sea relevante, se anima a los autores a depositar los datos en un repositorio público adecuado. Además, incluir una declaración de accesibilidad a los datos en su artículo y seguir los principios de citación de datos. Lo anterior está sujeto a consideraciones éticas, incluida la garantía de que el intercambio de datos no viole la protección de seres humanos u otras preocupaciones legales o de privacidad válida
Taylor and Francis Ltd	Se alienta a los autores a compartir y hacer que los datos estén abiertos cuando esto no viole la protección de sujetos humanos u otras preocupaciones válidas de privacidad del sujeto. Se anima a los autores a citar datos y proporcionar una declaración de disponibilidad de datos. Los autores deben hacer que sus datos estén disponibles libremente, bajo una licencia que permita la reutilización por parte de terceros para cualquier propósito legal. Los datos deberán ser fáciles de encontrar y totalmente accesibles. Además los datos deberán cumplir con los estándares FAIR (localizable, accesible, interoperable y reutilizable) según lo establecido en el área temática correspondiente.
NLM (Medline)	Los datos deben estar ampliamente disponibles de la manera más amplia posible en la medida en que sean compatibles con las leyes, regulaciones, reglas y políticas aplicables. Los datos deben ir acompañados de los metadatos y descriptores para que el intercambio sea significativo y útil para otros investigadores. Se ubicará en repositorios de datos con estándares comunes y una infraestructura establecida dedicada a la distribución adecuada de datos. Deben estar disponibles lo antes posible y durante el mayor tiempo posible.

El octavo atributo que queremos analizar tiene que ver con el contenido de los datos asociados al artículo. Venimos reflejando que en nuestra muestra, 15 artículos publicaron sus datos asociados en repositorios generales y temáticos, pero haciendo una verificación manual, de los artículos y los datasets vinculados, detectamos que sólo 11 son efectivamente datos de investigación, 3 tienen asociadas tablas con los resultados finales, 1 con referencias bibliográficas complementarias y 1 no está vinculado a los datos. En la tabla 10 se pone de manifiesto 9 se evidencia que solo una parte aplica el concepto de datos de investigación correctamente y el resto tiende a confundirlos con material complementario. Los autores declaran y depositan como conjunto de datos, materiales que no lo son y que podrían ir dentro del cuerpo de artículo. Se identificaron artículos donde se indica no tener datos asociados, pero al consultar el material complementario, se encontró el conjunto de datos disponibles

para descargar. En otros casos, hay artículos con datos asociados, pero al consultarlo no se trata de datos sino de figuras y tablas que deberían ir como material complementario.

Tabla 10: Contenido de los datos de investigación. Fuente: Elaboración propia.

Repositorio	URI	Contenido
Zenodo	http://doi.org/10.5281/zenodo.3701751	Datos estadísticos
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04326790	Datos sobre historias clínicas
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04326790	Datos sobre historias clínicas
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT03201185	Datos sobre historias clínicas
ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04321278	Datos sobre historias clínicas
Figshare	https://doi.org/10.1371/journal.pntd.0008065	Tablas con resultados finales
Figshare	https://doi.org/10.25384/SAGE.12294539.v1	Referencias bibliográficas complementarias
Figshare	https://doi.org/10.3389/fpubh.2020.00243.s001 https://doi.org/10.3389/fpubh.2020.00243.s002	Tablas con resultados finales
Figshare	https://doi.org/10.6084/m9.figshare.12850712.v1	Datos de la muestra
GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147863	Secuencia de ARN
GitHub	https://github.com/CSSEGISandData/COVID-19	Datos de la muestra
GitHub	https://github.com/DataLabUPO/CVOA_academic https://github.com/DataLabUPO/CVOA_LSTM	Código en Phyton y Java
Mendeley Data	http://dx.doi.org/10.17632/3xzm276pvk.1	Datos de la muestra
Mendeley Data	http://dx.doi.org/10.17632/98xcryjtnc.4	Datos de la muestra
Repositorio Digital de la UB	http://diposit.ub.edu/dspace/handle/2445/56611	Error de vinculación

El último atributo son las estadísticas, en algunos casos como se pueden identificar en la tabla 11 tienen métricas alternativas, por ejemplo Figshare además del número de visitas, descargas y citas permite mediante Altmetric tener acceso a las métricas sobre la atención en línea generada en redes sociales. Zenodo muestra estadísticas con el número de visitas, descargas y volumen de datos, vistas y descargar únicas. GitHub brinda estadísticas de citas y tiene un sistema de puntuación mediante estrellas. Mendeley Data muestra el número de citas, vistas y descargas. GEO y ClinicalTrials no especifican información al respecto.

Si relacionamos este atributo con el segundo atributo accesibilidad, nos muestra que la asignación del DOI a las citas aumenta la visibilidad de los datos. El primer paso para fomentar el uso de citas es incluirlas dentro del artículo asociado a los datos y añadir las referencias hacia los repositorios donde se encuentran depositados. Esta práctica ayuda a la obtención de estadísticas de citas, que se complementan con el número de vistas, descargas y métricas alternativas.

En el contexto de la pandemia, los resultados de las estadísticas nos llevan a cuestionarnos para futuras investigaciones si fueron son los datos o el artículo, los que tuvieron la mayor tasa de reutilización en investigaciones relacionadas con una solución efectiva para el virus.

Tabla 11: Datos estadísticos por repositorio. Fuente: Elaboración propia

Repositorio	Datos estadísticos
Zenodo	Estadísticas de vistas, descargas, volumen de datos, visitas y descargar únicas.
ClinicalTrials	No especifica
Figshare	Almetric + estadísticas de vistas, descargar y citas.
GEO	No especifica
GitHub	Estadísticas de citas y puntuación por estrellas
Mendeley Data	Estadísticas de citas, vistas y descargar
Repositorio Digital de la UB	Estadísticas de visitas y descargas por país y ciudad

El segundo grupo analizado fueron los datos disponibles a solicitud del autor (Fig. 11). Se hallaron 20 (6%) artículos que no tienen los datos disponibles. En 14 artículos los datos están disponibles a través de una solicitud al autor. En 2 casos el autor solicita adjuntar a la solicitud una propuesta metodológica sólida, con una descripción detallada de los objetivos del estudio y un plan de análisis estadístico para evaluar las solicitudes. En 2 artículos los datos no están disponibles por razones éticas con respecto a las leyes de privacidad de los participantes de la investigación, sin embargo el diccionario de datos del formulario de reporte del caso, el protocolo del estudio y la tabla de elementos del formulario de informes de casos están disponibles. 1 solo artículo tiene los permisos denegados para compartirlos y 1 afirma que todos los datos generados están incluidos en el contenido del artículo. Ningún caso tiene los datos bajo periodos de embargos.

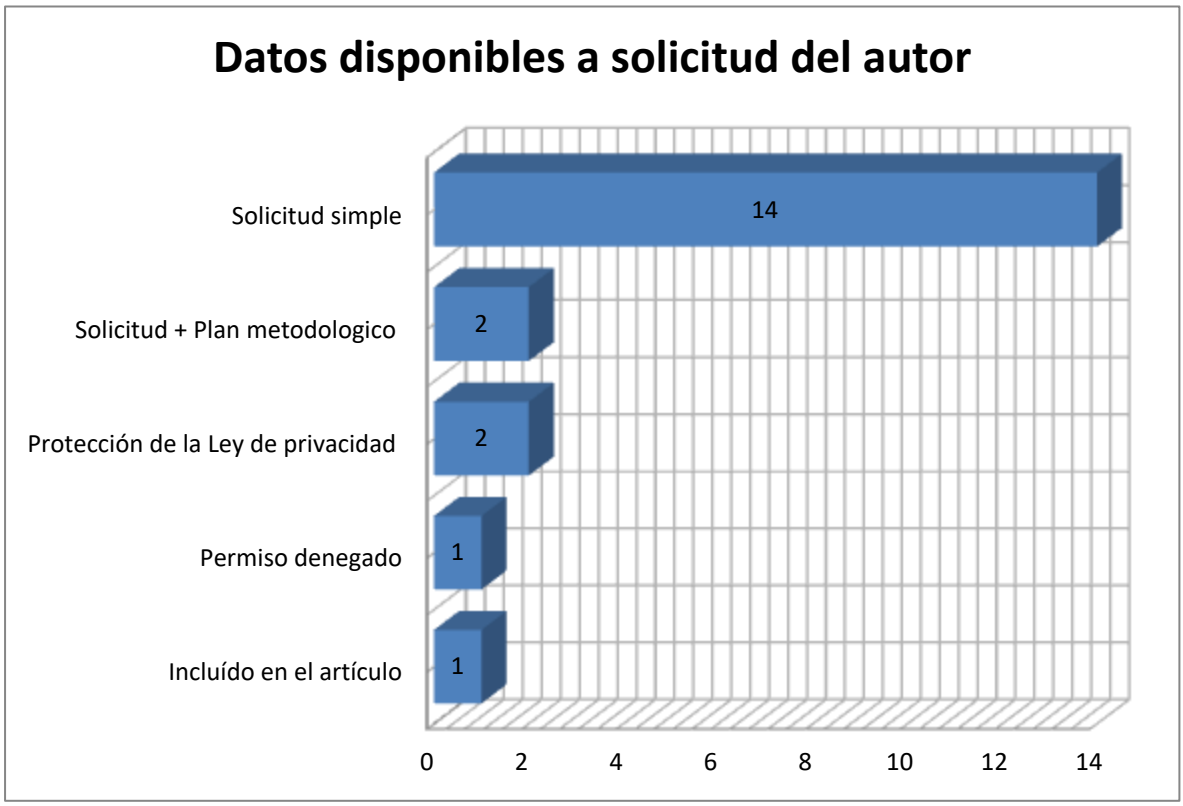

Figura 11: Datos disponibles a solicitud del autor. Fuente: Elaboración propia.

El tercer grupo analizado fueron los artículos con material complementario. De los 52 (72%) artículos con material complementario, se identificaron dos artículos que adjuntaron sus datos como material o información de soporte. Un artículo adjunto los ultrasonidos que se utilizaron para la evaluación de la atención en neonatos con COVID-19 y otro los datos para determinar los factores clínicos del mal pronóstico en pacientes con infección por COVID-19.

Reuniendo los tres grandes grupos de la muestra, el análisis puso de manifiesto que de los 87 artículos, solo 13 publicaron datos asociados: 11 en repositorios de datos generales y temáticos y 2 como material complementario del artículo, representando el 3.9% de la muestra analizada como se puede ver en la tabla 12.

Las razones que explican los resultados de nuestra meta-investigación implican a los agentes de la comunicación científica. La primera se relaciona con las editoriales y sus políticas de gestión de datos de investigación que sugieren, alientan e incentivan la publicación de manera voluntaria, pero no llega a ser un requisito obligatorio. Esto no significa que los datos deban estar en acceso abierto, recordemos que los principios FAIR indican que deben estar tan abierto como sea posible, tan cerrado como sea necesario. Los datos pueden tener permisos restringidos, pero sus metadatos pueden estar en acceso abierto, lo que ayudaría la reutilización para futuras investigaciones. La

segunda es el número reducido de editoriales y repositorios que brindan información completa y detallada sobre gestión de datos de investigación y la realización de un Plan de Gestión de Datos (DMP o PDG, según usemos sus siglas en inglés o en español). Los investigadores acuden en primera línea a las editoriales para ser asesorados, en algunos casos la editorial no tiene la información completa y actualizada o no cuenta con personal capacitado en el tema. La tercera son los incentivos. Los investigadores no reciben un incentivo adicional por la publicación de sus datos. De acuerdo al tipo de investigación los costos de depósito y preservación varían. Estos costos deben ser incluidos en el presupuesto inicial del proyecto, pero si no se elaboró o no se aplicó correctamente un PDG, se deben invertir dinero y tiempo adicional en la preparación de los datos para que puedan ser publicados.

Tabla 12: Contenido de los conjuntos de datos finales publicados de la muestra inicial. Fuente: Elaboración propia.

Nº	Repositorio y Revista	URI	Contenido
1	Zenodo	http://doi.org/10.5281/zenodo.3701751	Datos estadísticos
2	ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04326790	Datos sobre historias clínicas
3	ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04326790	Datos sobre historias clínicas
4	ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT03201185	Datos sobre historias clínicas
5	ClinicalTrials	https://clinicaltrials.gov/ct2/show/NCT04321278	Datos sobre historias clínicas
6	Figshare	https://doi.org/10.6084/m9.figshare.12850712.v1	Datos de la muestra
7	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147863	Secuencia de ARN
8	GitHub	https://github.com/CSSEGISandData/COVID-19	Datos de la muestra
9	GitHub	https://github.com/DataLabUPO/CVOA_academic https://github.com/DataLabUPO/CVOA_LSTM	Código en Phyton y Java
10	Mendeley Data	http://dx.doi.org/10.17632/3xzm276pvk.1	Datos de la muestra
11	Mendeley Data	http://dx.doi.org/10.17632/98xcryjtnc.4	Datos de la muestra
12	European Journal of Pediatrics	https://doi.org/10.1007/s00431-020-03706-4	Video de ecografías
13	PLoS ONE	https://doi.org/10.1371/journal.pone.0237960	Datos de la muestra

4. Conclusiones

Tras el estudio del contexto, las limitaciones metodológicas para el análisis de los conjuntos de datos vinculados a los artículos y el análisis detallado de la muestra de publicaciones españolas sobre COVID-19 y sus datos asociados, son varias las conclusiones que podemos establecer en este trabajo, que detallamos a continuación:

- Se evidencia la falta de indicadores para la evaluación de los datos de investigación, en los últimos años se ha hecho énfasis en crear normativas y herramientas digitales de comprobación y evaluación basados en la interoperabilidad y reutilización, dejando de lado los indicadores que describan los atributos y características para la evaluación cualitativas de los conjuntos de datos. Por otro lado, algunos repositorios de datos no incluyen la opción de exportación masiva de los registros para su análisis y evaluación, esto genera limitaciones en el desarrollo de la investigación.
- Las bases de datos de publicaciones científicas no cuentan con el campo "datos asociados", como uno de los campos de búsqueda y recuperación de sus búsquedas simples y avanzadas. La implementación de este campo disminuiría la confusión entre "datos asociados" y " material complementario".
- Las revistas solicitan, dentro de sus políticas de datos el depósito de manera voluntaria en repositorios de datos. Sin embargo, algunas revistas carecen de información clara que ayude a los autores a tomar decisiones sobre cómo y dónde depositar sus datos de investigación vinculados a las publicaciones.
- Existe confusión y ambigüedad sobre qué son datos de investigación y en qué formatos amigables deben compartirse. Se tiende a confundir que las tablas finales son los conjuntos de datos o las figuras de los datos lo son. Para asegurar la reutilización se deben publicar formatos libres como .csv, .xml, .txt y evitar publicarlos en .pdf y .doc
- Los resultados analizados respondieron a las preguntas iniciales de la investigación y se lograron los objetivos propuestos. Se identificó que los principales formatos utilizados por los investigadores para depositar sus datos son .pdf, . doc, .xls y .csv, bajo las licencias Creative Commons Atribución 4.0 Internacional (CC BY 4.0), con un DOI asignado que se utiliza para referenciar al artículo con sus datos. Para aquellos artículos que no tienen disponibles sus datos se deben solicitar al autor, ningún conjunto de datos de la muestra tenía periodo de embargo, para poder acceder se tiene que enviar una solicitud al autor pidiendo los datos, en algunos casos se debe adjuntar un plan metodológico. El análisis de todos los datos que formaron parte de la muestra fue la materia prima para elaborar una primera aproximación actual de los datos de investigación asociados a los artículos sobre la COVID-19.

En general, podemos decir que la pandemia producida por la COVID-19 ha provocado un aumento sin precedentes en el número de publicaciones, e induce también a un

cambio en la comunicación científica tradicional, propiciando el contexto necesario para que la Ciencia Abierta se difunda y se desarrolle. No obstante, aún falta mucho recorrido para que la publicación de datos asociados a una publicación sea la práctica habitual y aún más para que se sistematice su estudio a través de las bases de datos y repositorios implicados.

5. Bibliografía/Referencias

Abellán, A., Aceituno, P., Allende, A., Andrés, A., Bartumeus, F., Bastolla,U., Benavides, J., Cabal, B., Castillo, A., Chica, A., Cordech, M., Comas, I., Cuerdo, T., Damborenea, J., Domingo-Calap, P., Durán, M., Fernández, A., Fernández, C., Fernández, I., ... Vilà, M. (2020). *Una visión global de la pandemia COVID-19: Qué sabemos y qué estamos investigando desde el CSIC.* Consejo Superior de Investigaciones Científicas. Ministerio de Ciencia e Innovación. Gobierno de España. https://www.csic.es/sites/default/files/informe_cov19_pti_salud_global_csic_v2_1.pdf

Alcalá Ponce de León, M., & Anglada i de Ferrer, L. M. (2019). FAIR x FAIR. Requisitos factibles, alcanzables e implementables para un repositorio de datos de investigación FAIR. 43. https://www.recercat.cat/bitstream/handle/2072/356460/InformeFxF_maquetada_ESP.pdf

Amaro, B. (2020, abril 22). *Planeta biblioteca/ Entrevistada por Julio Alonso.* Universidad de Salamanca. http://radio.usal.es/programa/planeta-biblioteca/

Anglada, L., & Abadal, E. (2018). ¿Qué es la ciencia abierta? *Anuario ThinkEPI*, 12, 292. https://doi.org/10.3145/thinkepi.2018.43

Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are Scientific Data Repositories Coping with Research Data Publishing? *Data Science Journal*, 15, 6. https://doi.org/10.5334/dsj-2016-006

Australian National Data Service (s.f.). FAIR data training. https://www.ands.org.au/working-with-data/FAIRdata/training?utm_source=home-page-carousel&utm_medium=Web-banner&utm_campaign=FAIR-data-training

Bezjak, S., Clyburne-Sherin, A., Conzett, P., Fernandes, P., Görögh, E., Helbig, K., Kramer, B., Labastida, I., Niemeyer, K., Psomopoulos, F., Ross-Hellauer, T., Schneider, R., Tennant, J., Verbakel, E., Brinken, H., & Heller, L. (2018). *Open Science Training Handbook*. Zenodo. https://doi.org/10.5281/ZENODO.1212496

Brock, J. (2020, Abril 14). Rapid Registered Reports initiative aims to stop coronavirus researchers following false leads. *Nature Index*.

https://www.natureindex.com/news-blog/rapid-registered-report-coronavirus-aims-to-stop-researchers-following-false-research-leads

Caryn, R. (2020, Abril 16). Obesity Linked to Severe Coronavirus Disease, Especially for Younger Patients. *The New York Times*. https://www.nytimes.com/2020/04/16/health/coronavirus-obesity-higher-risk.html?smid=url-share

Comisión Económica para América Latina y el Caribe (18 diciembre de 2020). Gestión de datos de investigación. https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/tipos-datos

Comisión Europa (2020). Horizonte Europa. El próximo programa de inversión en investigación e innovación de la UE (2021-2027). https://ec.europa.eu/info/sites/info/files/research_and_innovation/strategy_on_research_and_innovation/presentations/horizon_europe_es_invertir_para_dar_forma_a_nuestro_futuro.pdf

Comisión Europea (2020). Una Estrategia Europea de Datos. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0066&from=EN

Consorcio de Servicios Universitarios de Cataluña (CSUC). (2020). Planes de Gestión de Datos. http://hdl.handle.net/2072/370510

Coronavirus pandemic: Nature's pledge to you. *Nature* 579, 471-472 (2020**)** https://doi.org/10.1038/d41586-020-00882-z

De Clercq, G.(2020, abril 8).Obesity is major COVID-19 risk factor, says French chief epidemiologist. *Reuters*. https://www.reuters.com/article/us-health-coronavirus-france-confinement/obesity-is-major-covid-19-risk-factor-says-french-chief-epidemiologist-idUSKBN21Q0S7

Digital Curation Center. (2013). Checklist for a Data Management Plan (v.4.0). http://www.dcc.ac.uk/resources/data-management-plans

Eisen, M. B., Akhmanova, A., Behrens, T. E., & Weigel, D. (2020). Publishing in the time of COVID-19. ELife, 9, e57162. https://doi.org/10.7554/eLife.57162

European Commission (2020). European Open Science Cloud (EOSC). https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/european-open-science-cloud-eosc_es

European Commission. (13 diciembre de 2019). Open Science. https://ec.europa.eu/info/sites/info/files/research_and_innovation/knowledge_publications_tools_and_data/documents/ec_rtd_factsheet-open-science_2019.pdf

European Commission. (2020). Data management. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

FAIRSFAIR: Fostering FAIR Data Practices in Europe (2020). Software and Tools. https://www.FAIRsFAIR.eu/tools-software

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, 22(2), 338-342. https://doi.org/10.1096/fj.07-9492LSF

Faundeen, J., Burley, T. E., Carlino, J. A., Govoni, D. L., Henkel, H. S., Holl, S. L., Hutchison, V. B., Martín, E., Montgomery, E. T., Ladino, C., Tessler, S., & Zolly, L. S. (2014). *The United States Geological Survey Science Data Lifecycle Model (USGS Numbered Series N.º 2013-1265; Open-File Report, p. 12).* U.S. Geological Survey. http://pubs.er.usgs.gov/publication/ofr20131265

Gagliardi, D., Cox, D., & Li, Y. (2015). Institutional Inertia and Barriers to the Adoption of Open Science. En E. Reale & E. Primeri (Eds.), The Transformation of University Institutional and Organizational Boundaries (pp. 107-133). SensePublishers. https://doi.org/10.1007/978-94-6300-178-6_6

Gómez, C. (02 de diciembre 2020). Pilar 1 – Ciencia Excelente. [Sesión de conferencia]. *Jornada de presentación en España del Programa Marco de Investigación e Innovación de la Unión Europea para el período 2021-2027 "Horizonte Europa"*, Madrid, España. https://cdn.eventscase.com/eventos.cdti.es/uploads/users/303506/uploads/890dd7a7ee673591371aea30490608b51d8b4795c4e17bd75b4759f2247ec9a47b908ee0b3046799fbe9315aa3dd21173842.5fd3454e2df04.pdf

Gøtzsche, P. C. (2012). Strengthening and Opening Up Health Research by Sharing Our Raw Data. *Circulation: Cardiovascular Quality and Outcomes*, 5(2), 236-237. https://doi.org/10.1161/CIRCOUTCOMES.112.965277

Hernández-Pérez, T., & García-Moreno, M.-A. (2013). Datos abiertos y repositorios de datos: Nuevo reto para los bibliotecarios. *El Profesional de la Información*, 22(3), 259-263. https://doi.org/10.3145/epi.2013.may.10

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, 3(1), 134-140. https://doi.org/10.2218/ijdc.v3i1.48

Konkiel, S. (2020). Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators. *Scholarly Assessment Reports*, 2(1), 13. https://doi.org/10.29024/sar.13

Kousouris, S. y Papadopoulos, G.(2020). Data Management (D9.7). Sustainable Energy Positive & Zero Carbon Communities (SPARCS).

https://www.sparcs.info/sites/default/files/2020-03/D9.7%20Data%20Management%20Plan_Final.pdf

Kwon, D. (2020). How swamped preprint servers are blocking bad coronavirus research. *Nature*, 581(7807), 130-131. https://doi.org/10.1038/d41586-020-01394-6

L'Hours, H., Stein, I. V., Huigen, F., Devaraju, A., Mokrane, M., Davidson, J., Vries, J. D., Herterich, P., Cepinskas, L., & Huber, R. (2020). CoreTrustSeal plus FAIR Overview. https://doi.org/10.5281/ZENODO.4003630

Larivière, V., Shu, F. & Cassidy, S. (5 de marzo de 2020). The Coronavirus (COVID-19) outbreak highlights serious deficiencies in scholarly communication. *LSE Impact Blog.* https://blogs.lse.ac.uk/impactofsocialsciences/2020/03/05/the-coronavirus-covid-19-outbreak-highlights-serious-deficiencies-in-scholarly-communication/

Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación. (2011). *Boletín Oficial del Estado*, 131, sec. I, de 02 de junio de 2011, 54387 a 54455. https://www.boe.es/boe/dias/2011/06/02/pdfs/BOE-A-2011-9617.pdf

López-Cózar, E. D., & Martín-Martín, A. (2020). *La viralidad de la ciencia defectuosa: El contagioso impacto mediático de un preprint en bioRxiv sobre el coronavirus y sus efectos en la comunicación científica.* https://doi.org/10.13140/RG.2.2.12666.44485

Méndez, Eva (2020, mayo 26). *Diálogo de Bibliotecas en Cuarentena/Entrevistada por Aída Montero*. Biblioteca Juan Bosch. https://www.youtube.com/watch?v=fm3GrE51Fdo

Ministerio de Ciencia e Innovación (2020). Estrategia Española de Ciencia, Tecnología e Innovación 2021-2027. https://www.ciencia.gob.es/stfls/MICINN/Ministerio/FICHEROS/EECTI-2021-2027.pdf

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37(1), 49-56. https://doi.org/10.3233/ISU-170824

OECD. (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. https://www.oecd-ilibrary.org/content/publication/9789264034020-en-fr

Petrilli, C. M., Jones, S. A., Yang, J., Rajagopalan, H., O'Donnell, L., Chernyak, Y., Tobin, K. A., Cerfolio, R. J., Francois, F., & Horwitz, L. I. (2020). Factors associated with hospitalization and critical illness among 4,103 patients with Covid-19 disease in New York City. *medRxiv*, 2020.04.08.20057794. https://doi.org/10.1101/2020.04.08.20057794

Pradhan, P., Pandey, A. K., Mishra, A., Gupta, P., Tripathi, P. K., Menon, M. B., Gomes, J., Vivekanandan, P., & Kundu, B. (2020). Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag. *bioRxiv*, 2020.01.30.927871. https://doi.org/10.1101/2020.01.30.927871

Research Libraries Group (2002). Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report. https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

Science in the time of COVID-19. *Nature Human Behaviour* 4**,** 327–328 (2020). https://doi.org/10.1038/s41562-020-0879-9

Scientific Data. (2020). Scientific Data recommended repositories (p. 98203 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.1434640.V16

Sixto-Costoya, A., Aleixandre-Benavent, R., Lucas-Domínguez, R., & Vidal-Infer, A. (2020). The Emergency Medicine Facing the Challenge of Open Science. *Data*, 5(2), 28. https://doi.org/10.3390/data5020028

Smale, N., Unsworth, K., Denyer, G., & Barr, D. (2018). The History, Advocacy and Efficacy of Data Management Plans. *bioRxiv*, 443499. https://doi.org/10.1101/443499

Strasser, C. (2012). Primer on Data Management: What you always wanted to know. https://doi.org/10.5060/D2251G48

Tibbo, H. y Jones, S. (2021). Research Data Management and Sharing. *Coursera*. https://www.coursera.org/learn/data-management

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

Travieso Rodríguez, C., & Ferreira Araújo, R. (2019). Aspectos metodológicos de los datos abiertos de investigación: Análisis de los conjuntos de datos de la colección SciELO incluidos en Figshare. *Revista española de Documentación Científica*, 42(3), 242. https://doi.org/10.3989/redc.2019.3.1597

Unesco (2019). La ciencia al servicio de la sociedad. https://es.unesco.org/themes/ciencia-al-servicio-sociedad

Unión Europea Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo, de 20 de junio de 2019 relativa a los datos abiertos y la reutilización de la información del sector público. Diario Oficial de la Unión Europea L172, de 20 de junio de 2019, pp. 56-83. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=uriserv:OJ.L_.2019.172.01.0056.01.SPA

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ*, 5, e3208. https://doi.org/10.7717/peerj.3208

Vitale, C. H., & Moulaison Sandy, H. (2019). Data Management Plans A Review. *DESIDOC Journal of Library & Information Technology*, 39(06), 322-328. https://doi.org/10.14429/djlit.39.06.15086

Wellcome Trust (2020, enero 31). Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. https://wellcome.ac.uk/coronavirus-covid-19/open-data

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wu, F., Wang, A., Liu, M., Wang, Q., Chen, J., Xia, S., Ling, Y., Zhang, Y., Xun, J., Lu, L., Jiang, S., Lu, H., Wen, Y., & Huang, J. (2020). Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. *medRxiv*, 2020.03.30.20047365. https://doi.org/10.1101/2020.03.30.20047365