# Utilización de recursos de Ciencia Abierta para la adquisición de información sobre artículos de divulgación

# *Open Science services and tools to get information from outreach articles*

**Javier Alonso del Saso**

alonsoj@ifca.unican.es

**Instituto de Física de Cantabria, (CSIC-UC)**

**Fernando Aguilar Gómez**

aguilarf@ifca.unican.es

**Instituto de Física de Cantabria, (CSIC-UC)**

## Resumen

La pandemia originada por el virus SARS-CoV-2 ha causado en la comunidad científica la necesidad de colaborar y promover prácticas de Ciencia Abierta, incluyendo el paradigma de datos abiertos para su reutilización. Este artículo explora un grupo de textos divulgativos publicados en la revista de The Conversation que, junto con diversas fuentes de información relacionadas con la Ciencia Abierta sirven para realizar un análisis de su contenido y su contexto, incluyendo información sobre sus autores, instituciones y disciplinas. Cruzando la información propia de los artículos con servicios abiertos como ORCID, Google Scholar o el uso de DOIs es posible dotar a los artículos de información contextual que se almacena en una base de datos. Esta información puede ser analizada utilizando técnicas de Ciencia de Datos para entender mejor todo el ciclo de vida de la investigación al mismo tiempo que facilita el descubrimiento de relaciones a nivel de artículo, temática o expertos.

## Palabras clave

Análisis de datos, Ciencia Abierta, Datos en abierto, FAIR, Localizable, Accesible, Reutilizable

## Abstract

The pandemic caused by the SARS-CoV-2 virus has caused in the scientific community the need to collaborate as well as promoting Open Science practices, including the open data paradigm to stimulate resing. In this article a set of outreach articles published in The Conversation are analyzed together with some Open Science services and tools to carry out an analysis of their content and context, including information about their authors, institutions and disciplines. This information can be analyzed to better understand the entire research life cycle while facilitating the discovery of relationships at the article, topics or experts.

## Keywords

Open Data, Open Science, FAIR, Data analytics, Findable, Accessible, Reusable

Recibido: 03/09/2022
Aceptado: 13/09/2022
DOI: https://dx.doi.org/IIMEI13-N25-014033
Descripción propuesta: Alonso del Saso, Javier; Aguilar Gómez, Fernando, 2022. Utilización de recursos de Ciencia Abierta para la adquisición de información sobre artículos de divulgación. *Métodos de Información*, **13**(25), 14-33

## 1. Introducción

Durante los últimos años hemos asistido a un incremento significativo de la disponibilidad de datos abiertos de las administraciones e instituciones públicas, promovido por una serie de leyes y directivas marco de la Unión Europea que refuerzan la importancia de la transparencia y la denominada democracia digital (van Keulen et al. 2019). Esto no se reduce únicamente a administraciones públicas, sino que, a nivel científico, está cobrando cada vez más relevancia y los investigadores son conscientes de las posibilidades que el acceso al conocimiento de forma abierta otorga para un desarrollo eficiente y ágil. El movimiento de ciencia abierta tiene por objetivo hacer la investigación científica y su difusión accesible a toda la sociedad, incluyendo los distintos componentes que hacen posible el desarrollo de un proyecto o experimento

científico, como los datos, métodos, software o las publicaciones. Dado que mucha de la actividad científica es financiada por fondos públicos, es legítimo hacer todo ese conocimiento accesible, por lo que grandes instituciones financiadoras como la Comisión Europea exigen adoptar el paradigma de la ciencia abierta dentro de sus proyectos. En el contexto del programa de financiación "Horizon Europe", se exige a los investigadores participantes en proyectos financiados adherirse a iniciativas como el European Open Science Cloud (EOSC) y adoptar buenas prácticas en ciencia abierta, en particular adoptando los denominados principios FAIR (Wilkinson et al., 2016), que propician que los datos producidos siguiendo ciertas características sean Encontrables, Accesibles, Interoperables y Reutilizables. Entre las buenas prácticas derivadas de esos principios, se encuentra el uso de tecnologías de la web semántica o web de datos, el uso de identificadores persistentes para identificar no sólo conjuntos de datos, sino personas involucradas en los proyectos y la utilización de metadatos que describan cada componente de la actividad científica con sus enlaces y relaciones. Todo ello puede explotarse tecnológicamente para tener un conocimiento detallado del contexto de un proyecto, en el que puedan conocerse cada uno de los elementos de su ciclo de vida y cómo se ha llegado a cierto resultado.

Sin embargo, para poder hacer que todo este conocimiento científico esté efectivamente al alcance de cualquier persona, es aún necesario la publicación de artículos de divulgación, que utilizan un lenguaje menos técnico y más sencillo de entender para colectivos no expertos en una temática científica particular. Aunque la mayoría de revistas de divulgación científica no requieren una información detalladas de todo el proceso científico, algunas sí que hacen uso de ciertos elementos de ciencia abierta, como identificadores persistentes para citar otros artículos (DOIs, Handle) o para identificar los investigadores involucrados en la publicación (ORCID). Toda esta "metainformación" unida al uso de agregadores de contenido científico como Google Scholar y el uso de APIs nos permite extraer información contextual de un artículo de divulgación y, a su vez, enriquecer en gran medida toda la información de la que disponemos, ayudándonos a entender el origen y el proceso de todo ese desarrollo de conocimiento.

Durante la pandemia de COVID-19 se ha demostrado que el paradigma de ciencia abierta estimula la colaboración y el avance del conocimiento (Van

Rossum & Drake, 2009) y su impacto se ha visto también reflejado en la publicación de artículos de divulgación que han tratado la pandemia desde distintos puntos de vista. El objetivo de este artículo es demostrar qué información contextual se puede extraer de un corpus de divulgación científica sobre COVID-19 a través de distintos métodos y tecnologías que permiten recopilar información sobre el contexto de las publicaciones. También la meta es proponer soluciones concretas para analizar toda esa información y visualizarla.

## 2. Metodología y fuentes de información

El presente artículo se ha desarrollado dentro de un proyecto de investigación donde se plantea el análisis del corpus completo de noticias sobre COVID19 publicadas por The Conversation durante el año 2020. Uno de los objetivos de este proyecto es obtener la información sobre especialistas en COVID19 en distintos ámbitos, algo factible explotando diversos recursos de ciencia abierta. También, se pretende estudiar el ciclo de vida de la información científica (Lenhardt et al., 2014), comenzando desde las fuentes originales citadas hasta su posible repercusión e impacto. De este modo, el trabajo se sustenta en distintas tecnologías de la información y de servicios abiertos disponibles a través de internet.

### Corpus The Conversation sobre COVID-19

La temática de COVID-19 ha sido durante esta época muy popular en múltiples campos de la ciencia, no sólo en los más directamente relacionados como la virología o la epidemiología, sino en otros donde la pandemia ha tenido cierto impacto. El mundo divulgativo-académico se encontró con muchos informes tratando la pandemia desde múltiples puntos de vista, especialmente en 2020, cuando existía una abundancia de preguntas para las que los investigadores buscaban respuesta desde distintos dominios. Esto nos ofrece un corpus divulgativo muy variado desarrollado por investigadores de múltiples campos que se enfoca en el mismo tópico en una revista que ofrece su contenido en libre acceso.

La colaboración con la revista The Conversation (The Conversation 2021) dentro del contexto del proyecto nos otorga acceso no solo al contenido en sí, sino a la organización de estos artículos, cuya estructura está mejor enfocada para el análisis automático del corpus incluyendo información contextual de los mismos. Además del contenido textual, se dispone de información sobre el número de comentarios que ha recibido el artículo, número de lecturas, título, dirección web, fecha de publicación y asignación temática por parte de la revista. Los editores también recogen información sobre los autores de la revista (i.e. nombre, género, institución a la que pertenecen, su campo de investigación, el identificador ORCID en el caso de incluirlo, número de artículos que ha escrito y número de comentarios en la revista). Asimismo, The Conversation guarda el nombre y localidad de las instituciones a las que pertenecen dichos autores.

Combinando la metainformación proporcionada por The Conversation con técnicas de análisis de documentos HTML y el uso de enlaces, relaciones e identificadores, además de servicios web disponibles de forma libre, podemos extraer gran cantidad de información contextual con el fin de analizarla para poder entender mejor el proceso de creación de nuevo conocimiento y sus relaciones con otros artículos, disciplinas de investigación, instituciones e investigadores.

## Información extraída del propio artículo

Para un análisis del contexto en el cual se ha escrito un artículo de divulgación conviene recoger toda la información accesible. Es por esto que además de aquella concedida por la revista, se complementa con información encontrada gracias a distintos servicios web de libre acceso. Uno de los datos recogidos ha sido la repercusión en redes sociales gracias a la funcionalidad de la revista, que permite compartir directamente un artículo. El número de veces compartido en las distintas redes sociales se publica en la página de cada artículo. Por otro lado, el proyecto se ha enfocado en la recuperación de datos recogidos de autores en los registros de ORCID y Google Scholar.

De forma adicional, los artículos en la página web citan o mencionan con enlaces web a sus recursos bibliográficos. Este dato nos ofrece una rica bibliografía para el corpus que podemos analizar para ofrecer más

información tanto de los artículos como de los autores, así como de trabajos previos relativos a la temática.

## Información adicional: Enriquecimiento del contexto

Los detalles del procedimiento y herramientas para la extracción de la información se mencionan en este apartado. El software para la extracción de los datos se ha desarrollado utilizando el lenguaje Python 3.9.4 (Van Rossum & Drake 2009), ya que posee una gran versatilidad y muchas herramientas bien documentadas que facilitan toda clase de tareas. Para este artículo nos centramos en las herramientas que proporciona para la comunicación web, lectura y transformación de datos además de formas de almacenaje de la información. También, para el análisis y almacenamiento de los datos se ha utilizado un servidor linux con 18.04.1-Ubuntu. Toda la información del proyecto ha sido volcada a una base de datos MySQL (Axmark & Widenius 2015). Esto permite interoperabilidad con distintas herramientas además de un centro o hub donde los miembros del proyecto pueden acceder a la última versión de la información del proyecto.

## Comunicación web

Todos los recursos del proyecto son originarios de la web. Por eso, es importante tener buenas herramientas y aprovechar los recursos a disposición para conseguir la información.

Gran parte de la comunicación en Internet se basa en el protocolo HTTP(S), donde gran variedad de documentos se comparten continuamente. Dependiendo del servidor y del servicio disponible existen distintos formatos para el intercambio de datos. Los más habituales son:

- HTML es el formato más popular y en el cual se basan muchas páginas web.
- XML es un formato basado en etiquetas (al igual que HTML). Está diseñado para estructurar los datos y la información.
- JSON usa contenedores con nombres para estructurar los datos.

Todos los formatos anteriores son de uso abierto y preparados para ser accesibles por cualquier tipo de tecnología.

Se hace uso de APIs (Application Program Interfaces), interfaces que permiten el intercambio automatizado de información con un recurso web. Los formatos y protocolos anteriores citados se emplean para transportar la información. En el caso de las APIS web, es necesario que estén documentadas puesto que utilizan un lenguaje propio para navegar por los recursos.

ORCID es un ejemplo de servicio que recoge información de científicos, enlazado con artículos y méritos y disponible a través de una plataforma web o mediante el uso de una API.

Google Scholar es otro servicio, pero no ofrece una API abierta, por lo que es necesario recuperar la información en formato HTML. Al no estar la información organizada es necesario conocer bien la estructura para poder extraer y depurar la información, pero podemos aprovechar ciertas herramientas existentes que nos facilitan este proceso.

## 3. Análisis

En este apartado se describen los procedimientos requeridos para recoger, depurar y coleccionar la información disponible en la web que nos proporciona datos sobre el contexto de los artículos, como publicaciones relacionadas, temáticas, información sobre los autores o instituciones, etc. Para recoger la información nos basamos en los datos que se han reunido de los autores y de sus publicaciones en la revista The Conversation.

En la recolección nos basamos en la incorporación de datos sobre las métricas de los artículos divulgativos, su información bibliográfica, así como la extracción de información relevante de los autores de otras fuentes ajenas a la revista, como en este caso, de ORCID y Google Scholar.

### Métricas de los artículos

Para un análisis de las métricas de los artículos nos basamos en la información que nos proporciona la propia revista. Al no encontrarnos con otras herramientas presentes en revistas académicas, nos limitamos a extraer los datos del número de tweets, facebook shares (FS) y linkedin shares (LS). Esta información la encontramos dentro del mismo recurso web del artículo. Sin embargo, a la hora de automatizar el proceso, es recomendable utilizar herramientas y recursos que estén diseñados para dicha metodología.

La comunicación utiliza el formato json para codificar la información, que es menos visual que el formato HTML (comúnmente más usado para comunicaciones HTTP(S)) pero fácilmente legible para una máquina o software. Al completar la dirección web con el identificador de un artículo podemos extraer la información sin dificultad. Esto se puede realizar con peticiones a través del protocolo HTTP utilizando el lenguaje Python.

Una vez conseguida las métricas, introducimos tres nuevos campos que recojan cada una de las métricas como valores enteros positivos dentro de la tabla artículos.

## Registro de autores en ORCID

El registro de ORCID nos permite hacer una referencia cruzada con la información original que la revista nos ofrece de los autores. De esta manera podemos reunir más información para mejor comprender y perfilar a dichos investigadores.

Gracias a que The Conversation nos ofrece el identificador de ORCID del autor (en el caso de que posea uno), la inferencia se vuelve directa.

ORCID proporciona libremente una API (Application Program Interface) para hacer llamadas automáticas a los servidores. A través del punto de acceso habilitado, es posible solicitar información de autores con un identificador ORCID.

La API devuelve objetos XML que contienen varios campos principales que a su vez se dividen en subcampos y así sucesivamente. El campo denominado RECORD contiene toda la información relevante del autor. Gracias al

servicio habilitado (ORCID TEAM, 2020), se puede obtener información a través de consultas al campo mencionado.

Los campos recogidos por ORCID son los siguientes: Palabras claves, distinciones, estudios, empleos, formación, fondos económicos, invitado a algún evento, afiliaciones, revisiones de artículos, cualidades, recursos (i.e. artículos de prensa y científicos), servicios, trabajos y biografía.

La gran mayoría de estos campos son completados por el propio autor y es él quien debe actualizarlos con la última información. Esto puede resultar ser un problema, ya que en ocasiones estos campos pueden permanecer vacíos. Además, no existe garantía de que la información recogida sea actual.

## Registro de autores en Google Scholar

Google Scholar es una plataforma web que recoge información sobre artículos y libros científicos. Además, también recoge datos de autores, dependiendo de si tienen una cuenta registrada dentro de esta plataforma. Esta herramienta puede proporcionar información sobre artículos publicados por un autor, métricas de los autores y de sus artículos, así como métricas de impacto.

Al contrario que ORCID, Google no proporciona una API o herramienta automática para recoger o analizar la información. Es por esta razón que han surgido herramientas que imitan a una API mediante comunicación HTTP utilizando librerías y software. En concreto, Scholarly (Cholewiak et al. 2021) es una librería para el lenguaje Python que nos permite realizar llamadas a Google Scholar y ordenar la información como si la respuesta fuera en formato XML.

Uno de los mayores problemas es la comunicación entre la librería y el servicio web. Al utilizar una herramienta que automatiza la tarea, la comunicación con el servidor se vuelve muy eficiente y el número de peticiones por segundo pueden incrementar de forma considerable. Es por esto que Google, como muchas otras compañías que ofrecen servicios online, tiene una estructura preparada para bloquear aquellas máquinas que ocupen gran parte de la capacidad de procesado o de ancho de banda en la web. En

realidad, basta que la herramienta sea detectada como no-humano como para bloquear la señal.

Es muy importante tener medidas dentro del código para saber responder a la hora de un fallo de conexión, como, por ejemplo, mantener guardados los datos que sí han podido ser salvados antes de la desconexión, preparar el resumen de la comunicación una vez que el problema es solventado o evitar realizar muchas llamadas en un período corto de tiempo para mantener una conexión estable.

Al no existir identificadores para autores registrados en Google Scholar, y la información recogida en The Conversation de los autores no incluye ningún recurso web que apunte a ellos, debemos realizar una búsqueda de nombres como identificadores de los autores. Esto presenta varias dificultades.

El nombre como identificador implica que existe un identificador para varias personas con el mismo nombre. Aunque la probabilidad de dos coincidencias exactas es menos probable debido al uso de doble apellido para autores en español, tampoco ofrece garantía de que los registros recogidos sean de los mismos autores. La mejor opción es recoger el mejor resultado que ofrece Google (i.e. máxima coincidencia) cuando se busca el nombre de un autor.

Por otro lado, se puede producir que los nombres estén contraídos o existan erratas en ellos (e.g. las iniciales del nombre, o que falte el segundo apellido). Por esta razón también se realizará una búsqueda con aquellos autores cuyos nombres no han sido coincidentes pero muy similares esta búsqueda se hace de forma automática y manual para albergar el mayor número de autores posibles.

Los campos más relevantes de autores registrados en Google Scholar son: Afiliación, número total de citas, número de citas de los últimos 5 años, número de citas anuales, dominio de email, índice H, índice H de los últimos 5 años, índice i10, índice i10 de los últimos 5 años, campos relacionados y publicaciones. Dentro de las publicaciones se registra el número de citas, año de publicación, y título.

### Estructura de artículos y autores

Todos los datos del proyecto son volcados en una base de datos MySQL. Los datos recogidos se estructuran para facilitar el análisis, manipulación, almacenaje además de la detección de fallos y errores. En la base de datos, la información se organiza en tablas. Existen 2 tablas principales, una para autores y otra para artículos publicados en The Conversation.

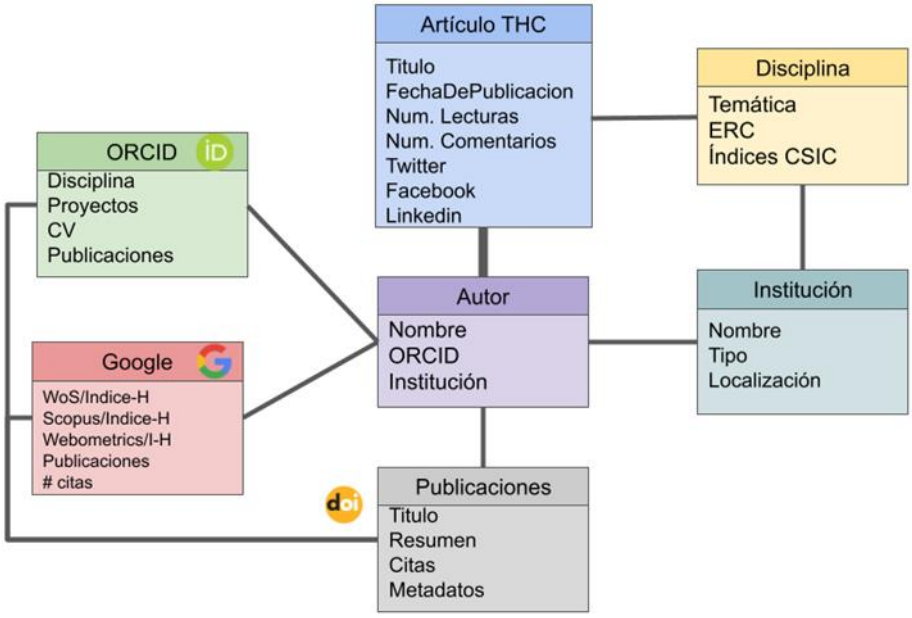


***Figura 1.*** *Organización del modelo de datos en tablas de los datos recogidos para el proyecto, recogiendo los datos de los artículos de The Conversation (THC) categorizados según las temáticas descritas por Índices CSIC y European Research Council (ERC).*

## Métodos de representación de la información

Al encontrarse un volumen considerable, presentamos los métodos para desarrollar algunas de las visualizaciones diseñadas para ofrecer la mayor cantidad de información sobre el conjunto de datos recogidos de autores y artículos. De este modo se puede sintetizar y resumir los resultados de un análisis de forma eficiente.

El análisis de correspondencia múltiple (MCA) estudia la similitud entre elementos agrupados por distintas clasificaciones (Abdi & Valentin, 2007). Permite definir una representación reducida (e.g. una figura 2-dimensional) de las similitudes como medidas de distancia relativas entre ellas basado en la coincidencia de agrupación en distintas clasificaciones de dos elementos. Este

método produce a su vez una representación similar para las mismas agrupaciones y establecer una relación de similitud entre ellas.

Al tener varias clasificaciones de los artículos según su contenido podemos hallar una representación que nos muestre si existe algún tipo de estructura subyacente tanto de los artículos como de las agrupaciones. Por otro lado, guardamos datos numéricos relacionados con número de publicaciones científicas del autor o comentarios en la revista, entre otros. Como no existe una relación entre las variables clara de la que podamos partir, debemos centrarnos en descifrar si existe algún tipo de relación en primer lugar. Para esta tarea empleamos el coeficiente de la correlación de Spearman (Dodge, 2008), la cual determina si existe para una variable independiente y otra independiente, una relación de monoteneidad, por ejemplo, cuando una variable varíe, la otra lo hace en la misma dirección (no necesariamente siguiendo una ley de proporcionalidad directa). El coeficiente puede tomar valores entre -1 y 1, siendo sus valores extremos una relación perfecta de monoteneidad.

## 4. Resultados

A partir de la información extraída con los métodos descritos es posible hacer un análisis que nos permita entender el contexto en el que se ha desarrollado un trabajo, así como relaciones entre diferentes elementos de la investigación. A continuación, presentamos los resultados obtenidos a partir del análisis de la combinación de información de autores y artículos de la revista con la recogida de las fuentes de ORCID y Google Scholar.

### Estructura de los datos

Con el fin de mantener los datos en una estructura lógica para evitar errores, así como para asegurar su conservación y facilitar las operaciones, se ha utilizado un sistema gestor de bases de datos. Éste alberga la información basándose en una estructura de tablas relacionales con los campos más relevantes recogidos en la Figura 1.

Las tablas principales siendo Articulo THC y Autor que recogen la información de los artículos y autores respectivamente. Otras tablas como Disciplina y Institución proporcionan información adicional categorizada de tal forma que evita erratas y fuerza dicha clasificación sobre los datos. La primera tabla resume la información sobre la categorización de los artículos de las distintas fuentes. La segunda recoge la institución (e.g. centro, instituto, laboratorio, universidad) a la que pertenece cada autor según apunta The Conversation.

## Fuentes en abierto sobre los autores

Desde los datos disponibles en abierto, se ha rescatado diversa información sobre los autores. Una de las fuentes es ORCID, que organiza la información en apartados o campos y cada uno de ellos debe ser rellenado por el autor, institución o revista (dependiendo del caso) de forma manual o automática. Por esto, no todos los campos son rellenados y aquellos que lo están, varían de autor en autor. La Figura 2 muestra un listado de los campos ordenados según el número de autores que tienen información en dicho campo. Encontramos que los campos más populares son Trabajos y Empleos, que corresponden a las publicaciones en revistas académicas por un lado (generado muchas veces de forma automática a partir de DOIs) y a la lista de empleo (e.g. catedrático, doctorando). También puede aparecer información de la organización y departamento.

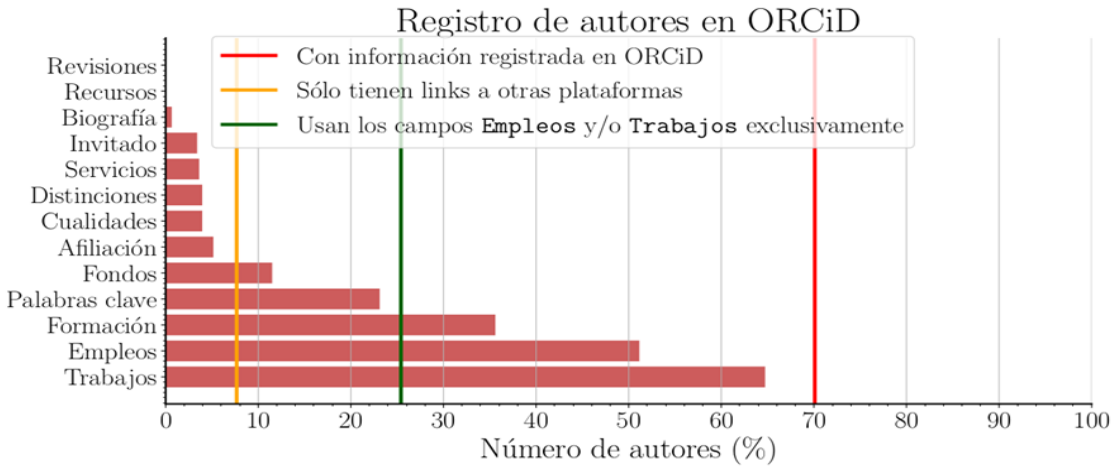


***Figura 2.*** *Porcentaje de autores registrados en ORCID que tienen cada campo rellenado. Las líneas verticales corresponden al porcentaje de autores que sólo recogen links a recursos externos a ORCID (izq.) y autores que están registrados en ORCID (der.). El eje horizontal corresponde al número de autores registrados en la revista.*

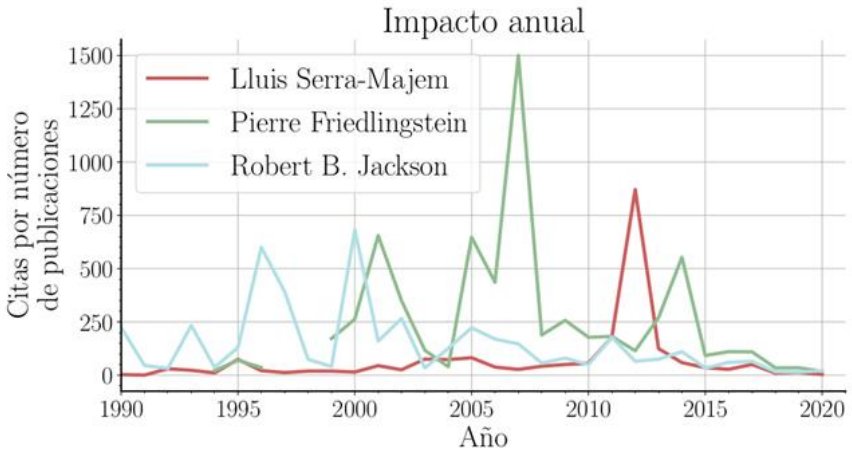


***Figura 3.*** *Visualización del impacto de los autores más relevantes. Evolución del índice de citas por publicación del autor.*

Encontramos que el 70.10 % de los autores tienen asociado una cuenta de ORCID, 7.67 % lo usan para apuntar a otros recursos web y 25.41 % usan los campos de Trabajos y Empleos exclusivamente. Esta información es adicional a la que puede proporcionar Google Scholar con respecto al impacto, que podemos ver en la Figura 3. Para una selección de tres de los autores con más relevancia y con publicaciones en The Conversation podemos ver el número de citas que reciben anualmente para sus artículos. Asimismo, a partir de este servicio podemos recopilar el índice H de los citados autores, que podemos contemplar en la Figura 4. Los autores han sido elegidos como muestra representativa de entre los 913 autores que han publicado en The Conversation sobre COVID-19 en 2020, por ser los que mayores citas poseen según los servicios de Google Scholar. De acuerdo a la información recogida, encontramos que Lluis Serra-Majem tiene 1399 artículos con un total de 97396 citas, Robert B. Jackson tiene 992 artículos con 95208 citas y Pierre Friedlingstein 425 artículos con 100438 citas.

Los picos máximos tanto en el impacto de cada autor (años 2000, 2007,2012) como en su índice-H (años 2011, 2013, 2015) no coinciden. Sin embargo, aclara la repercusión de un autor, observado desde distintos puntos de vista, algo que es esencial para reforzar la importancia de la ciencia abierta en un nuevo paradigma.

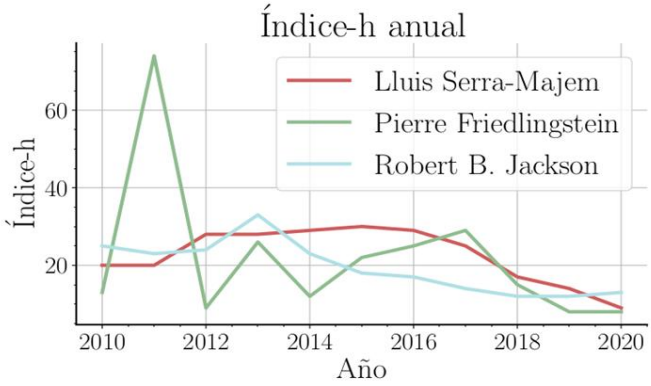


***Figura 4.*** *Evolución temporal del índice-h anual de los autores con mayor impacto de la muestra.*

Otra opción de análisis a partir de la información recopilada es el estudio de las referencias de los artículos de divulgación. En particular, gracias al protocolo HTTP y por cómo los servidores almacenan la información, podemos observar las fechas de creación o modificación de un recurso. En la Figura 5 se recogen las referencias bibliográficas empleadas en la muestra de artículos de The Conversation haciendo uso de la fecha de última modificación de estos recursos. Podemos observar que la gran mayoría (92.0 %) de recursos han sido modificados en el año 2021. Recordemos que las publicaciones de los artículos de muestra son del año 2020. Esto se debe a que la fecha recogida comprende al recurso web y no a su contenido (e.g. cambios si aparecen nuevos comentarios). Sin embargo, podemos observar recursos con fecha de modificación a finales de los años 90 y vemos que el número de referencias suele ir en aumento. Para un tema novedoso como el de la pandemia de COVID-19, es lógico pensar que los artículos hagan referencias a artículos próximos en el tiempo.

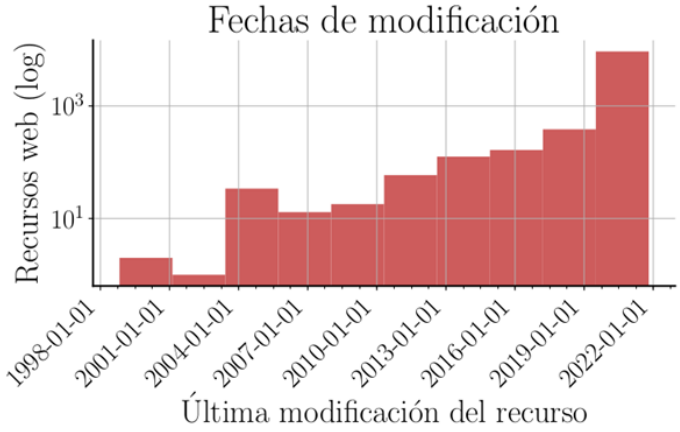


***Figura 5.*** *Histograma de bibliografía de artículos ordenados cronológicamente en escala logarítmica.*

La relación que guardan los datos de Google Scholar y los datos de la revista se manifiestan en la Figura 6. Encontramos relaciones muy grandes debido a la derivación de una variable con respecto a la otra. Estas son: índice-h con el número de citas; número de archivos HTML con su porcentaje; comentarios por artículo y número de comentarios.

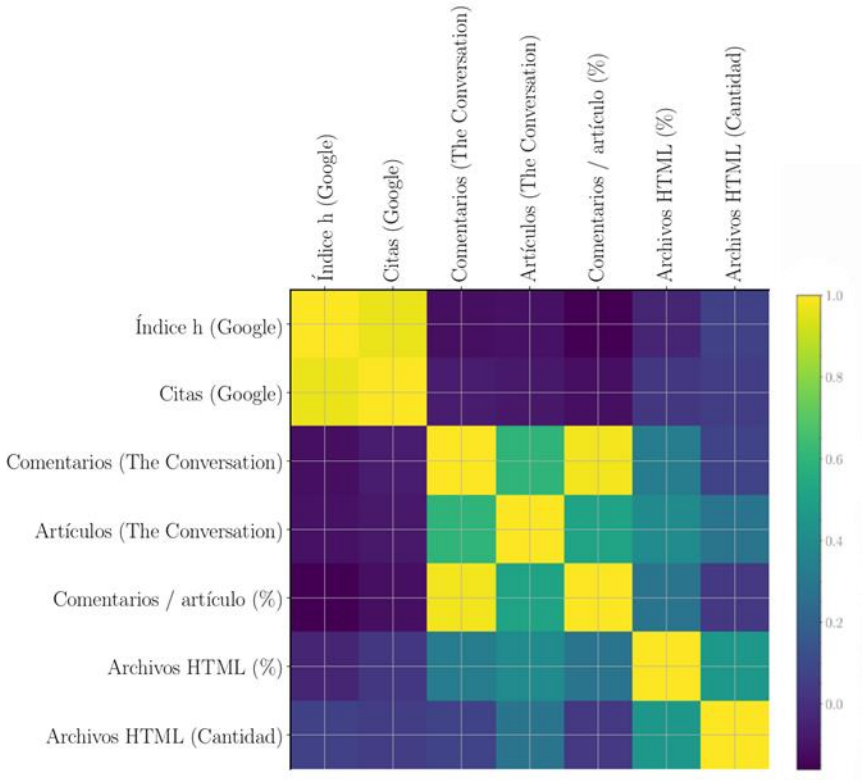


***Figura 6.*** *Correlación de Spearman de distintas variables recogidas de los autores en la revista, Google Scholar y de los recursos bibliográficos encontrados en los artículos.*

Sin embargo, existen otras relaciones interesantes, el número de artículos y el número de comentarios y el porcentaje de recursos HTML con el número de artículos. La primera indica que la actividad del autor en la revista se puede ver reflejado tanto en su capacidad de publicación como en su capacidad de comentar. La segunda relación se debe a que el autor que contiene más artículos, también tiene una bibliografía más extensa (especialmente en formato HTML).

El resumen visual de nuestro análisis ejemplar lo podemos encontrar en la Figura 7. Podemos distinguir distintos perfiles de densidades para distintas variables. Encontramos que utilizando el método de MCA (Abdi & Valentin, 2007) un gran número de categorías se encuentran en la zona central para

clasifGoogle, clasifOrcid, lo cual viene a demostrar que existe una relación subyacente en las distintas categorías definidas por las dos plataformas.

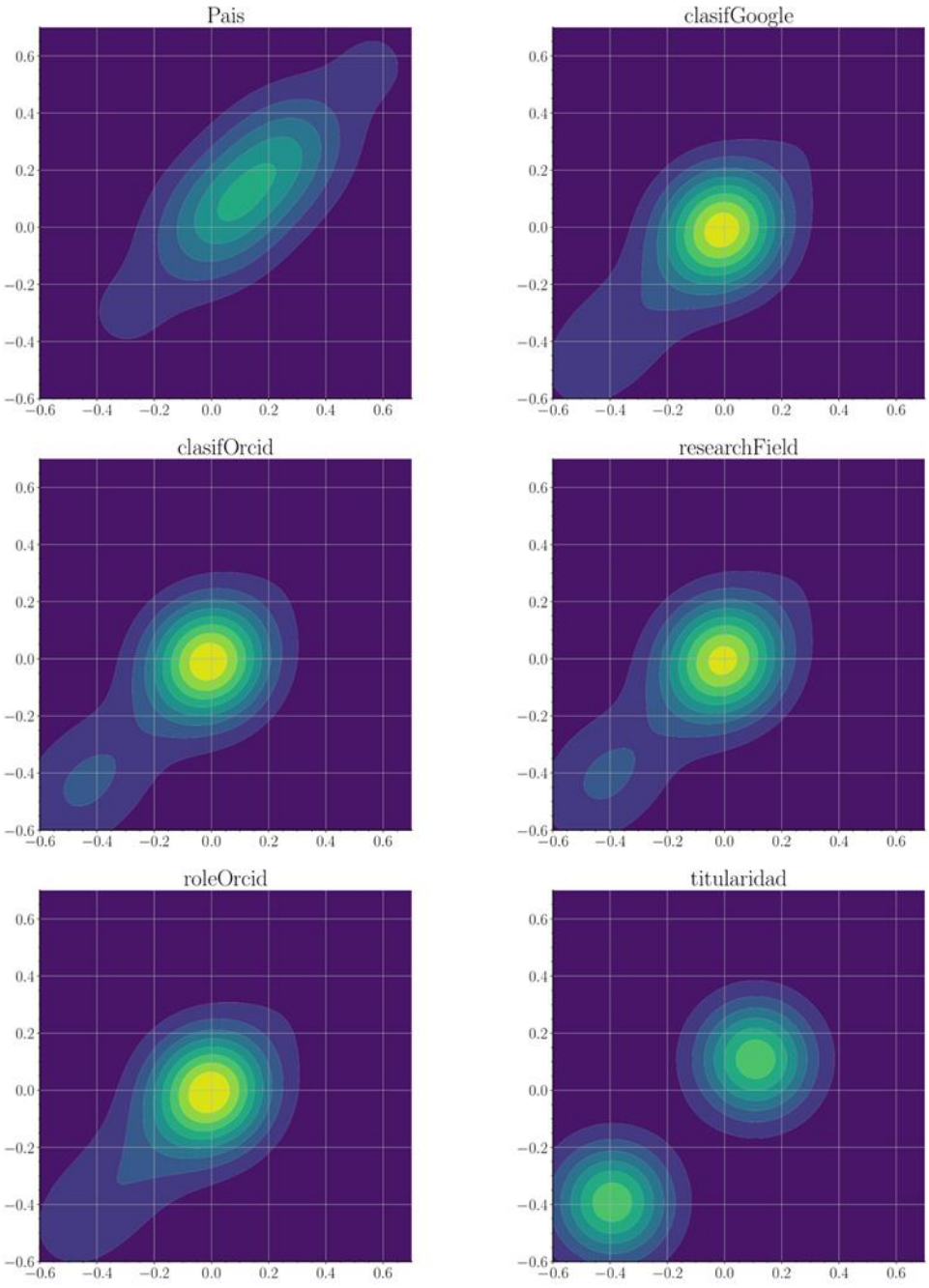


***Figura 7.*** *Densidad de categorías de cada variable en el espacio reducido del método MCA. Los colores más claros reflejan dónde se encuentra mayor densidad.*

## 5. Conclusiones

La ciencia abierta está cobrando cada vez más importancia, siendo una nueva forma de acercar el desarrollo del conocimiento a la sociedad de forma

transparente y accesible. Incluye nuevos elementos que tradicionalmente no han estado presentes en la forma en la que se expresaban los desarrollos y hallazgos científicos. En lugar de limitarse a artículos científicos y libros, se incorporan otros componentes del flujo científico, como datos, software, servicios e incluso publicaciones divulgativas. Además, el acceso universal a internet ha modificado la forma a la que se accedía a la información científica, haciéndola mucho más fácil y accesible. Pero no sólo eso, facilita también la discusión, la interacción y el enriquecimiento de la información a través de las redes sociales. La forma en la que hacemos ciencia ha cambiado, por lo que debemos cambiar también la forma en la que medimos su impacto.

Este artículo expone diversos métodos para la recopilación automática de información relativa a una publicación de divulgación científica. A modo de ejemplo, se han utilizado recursos abiertos y accesibles que permiten la recopilación de estos datos de forma sencilla, enriqueciendo en gran medida el contexto de la publicación. Aunque está centrado en una lista cerrada de artículos de divulgación sobre COVID-19, estos métodos son extrapolables a cualquier corpus con cierta información se partida, como el nombre de los autores, categorías del artículo o instituciones a las que pertenecen. En publicaciones científicas tradicionales esto suele ser más sencillo, ya que adoptan el uso de metadatos con esquemas estándares e información regular. Además el uso de identificadores persistentes identifica unívocamente distintos tipos de recursos, que a su vez aportan nueva información. Esto es totalmente esencial para la adopción de los principios FAIR. En publicaciones de divulgación, estas buenas prácticas están parcialmente adoptadas, como por ejemplo el uso de identificadores para autores (ORCID). Estos sistemas estándar contribuyen a facilitar el acceso a información homogeneizada, por lo que sería interesante que revistas como The Conversation incorporaran más de estas funcionalidades.

Se ha extraído información no sólo de los artículos en sí, como sus posibles categorías, disciplinas, referencias, lecturas, repercusión en redes sociales, etc. sino datos relativos a los autores, métricas de impacto de los mismos, instituciones, posibles colaboraciones e infinidad de conocimiento que podría extenderse para establecer relaciones entre los distintos actores involucrados y sus interacciones con otros colegas. De este modo, ha sido posible recopilar información sobre especialistas en COVID-19 diferenciados por disciplinas,

así como su relación con distintos temas. Además, hemos enriquecido la visión disponible de los artículos para entender de una forma más completa todo el ciclo de vida de la investigación, trabajos previos, herramientas y recursos utilizados, etc. Por último, se pueden identificar relaciones entre las distintas formas de categorizar artículos, dominios de los autores e instituciones, descubriendo patrones que no se distinguen a simple vista.

Con este tipo de técnicas y su análisis somos capaces de entender mejor cómo se ha llegado a una conclusión o resultado científico, alcanzando así el objetivo del proyecto de contextualizar los artículos para obtener una red de expertos y entender mejor el ciclo de vida de la investigación. Esto aumenta la transparencia, democratizando el acceso a la ciencia y asegurando en gran medida la reproducibilidad de los hallazgos científicos.

## 6. Agradecimientos

Este trabajo se ha realizado gracias a la colaboración dentro de la Plataforma Temática Interdisciplinar "EsCiencia", que promueve el uso del español como lengua de comunicación científica. En particular, dentro del proyecto "COVID 19 en español: Investigación interdisciplinar sobre terminología, temáticas y comunicación de la ciencia", financiado como proyecto intramural del CSIC y que se plantea el análisis del corpus completo de noticias sobre COVID19 publicadas por The Conversation (en español).

Nuestro más sincero agradecimiento a todos los miembros del equipo del proyecto, en particular a Elea Giménez, José Ignacio Vidal, César González-Pérez, Ana García y Teresa Abejón.

## 7. Bibliografía

ABDI, H., & VALENTIN, D., 2007. Multiple correspondence analysis. *Encyclopedia of measurement and statistics*, **2**(4), 651-657.

AXMARK, D., & WIDENIUS, M., 2015. *MySQL 5.7 reference manual*. Redwood Shores, CA: Oracle. http://dev.mysql.com/doc/refman/5.7/en/index.html

BESANÇON, L., PEIFFER-SMADJA, N., & SEGALAS, C., 2021. Open science saves lives: lessons from the COVID-19 pandemic. *BMC Medical Research Methodology*, **21**(117). doi: 10.1186/s12874-021-01304-y

CHOLEWIAK, S., IPEIROTIS, P., SILVA, V., & KANNAWADI, A., 2021. *scholarly*. Zenodo. doi: 10.5281/zenodo.5765779

*THE CONVERSATION*, 2021. The Conversation: Noticias, investigaciones, ideas y análisis en profundidad firmados por académicos e investigadores de primera línea. [Consulta Diciembre 9, 2021]. Disponible en: https://theconversation.com/es

DODGE, Y. 2008. *Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics*. Springer. doi: 0.1007/978-0-387-32833-1

LENHARDT, W., AHALT, S., BLANTON, B., CHRISTOPHERSON, et al., 2014. Data management lifecycle and software lifecycle management in the context of conducting science. *Journal of Open Research Software*, **2**(1). http://doi.org/10.5334/jors.ax

ORCID TEAM, 2020. *ORCID API v3.0 Guide*. GitHub. [Consulta febrero 9, 2022]. Disponible en: https://github.com/ORCID/orcid-model/tree/master/src/main/resources/record_3.0#orcid-api-v30-guide

VAN KEULEN, I., KORTHAGEN, I., NIELSEN, R. Ø., et al,. 2019. *European E-Democracy in Practice*. Springer International Publishing. ISBN: 978-3-030-27184-8

VAN ROSSUM, G., & DRAKE, F. L. 2009. *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. CreateSpace Independent Publishing Platform. ISBN: 978-1-4414-1269-0

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON,, G., AXTON, M., & BAAK, A., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, **3**(160018). https://doi.org/10.1038/sdata.2016.18.