

SESIÓN 4

Indicadores de Ciencia Abierta

Contenidos de la sesión

1 ¿Qué es la Ciencia Abierta?	1
El concepto de Ciencia Abierta	1
Los grupos de Expertos de la UE	2
2 Acceso Abierto	3
Orígenes	3
Acceso Abierto y bibliometría	4
Unpaywall: el estándar actual	6
3 Los datos de investigación	8
Data sharing e impacto científico	8
Fuentes de datos de investigación: Data Citation Index y DataCite	9

1 ¿Qué es la Ciencia Abierta?

El concepto de Ciencia Abierta

El término Ciencia Abierta se puede rastrear a 2015, cuando el entonces comisario de Investigación, Ciencia e Innovación de la Comisión Europea [lo menciona por primera vez](#).

I am convinced that excellent science is the foundation of future prosperity, and that openness is the key to excellence. We are often told that it takes many decades for scientific breakthroughs to find commercial application.

En este discurso no se da una definición exacta del concepto, aunque sí que se relaciona con otros dos conceptos: **Ciencia abierta e integridad en la investigación**.

Otro documento que surge ese mismo año y de manera simultánea es el **Metric Tide**, un informe de expertos que pretende informar sobre el uso de indicadores bibliométricos para la evaluación en el

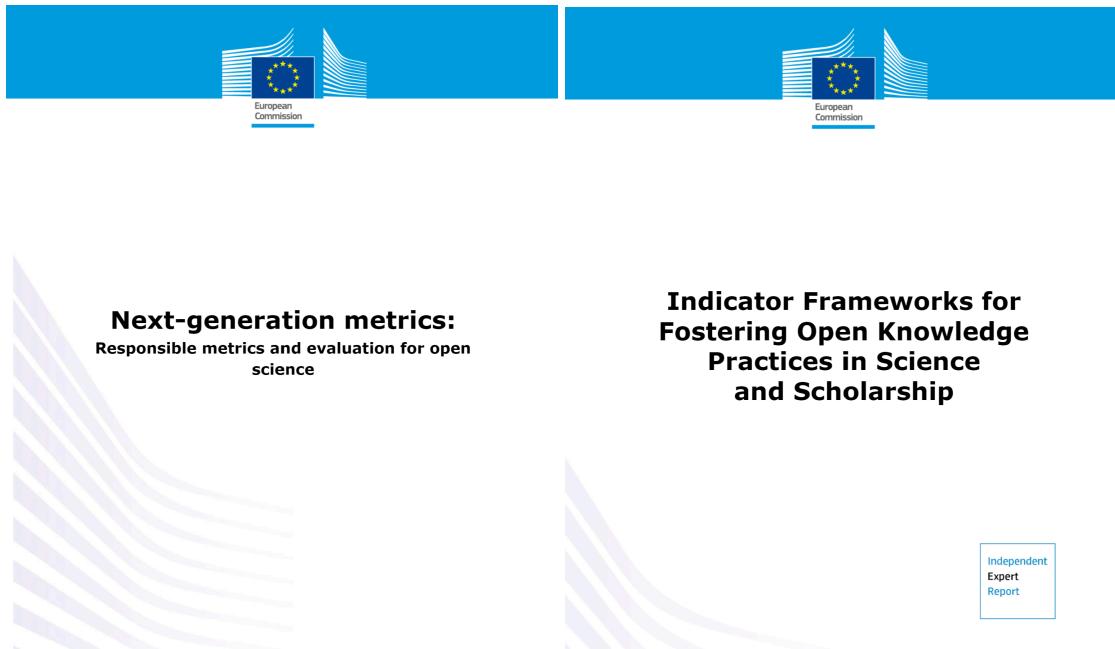


Reino Unido.

Allí que se habla por primera vez de **métricas responsables**, también se habla de **transparencia y compartición de datos en abierto**.

Los grupos de Expertos de la UE

La falta de indefinición hace que sea difícil decidirse a medir la Ciencia Abierta y empiezan los esfuerzos por definir y operacionalizar.



El resultado es un desastre. Los diferentes grupos de expertos no son capaces de ponerse de acuerdo, ya no en qué indicadores emplear, sino en una [definición operacional](#) del término.

Aunque sí acuerdan una serie de dimensiones y/o conceptos relacionadas con la Ciencia Abierta:

- Ciencia ciudadana
- **Data sharing** 🤝
- **Acceso Abierto** 🤝
- Impacto social
- Reproducibilidad
- Transparencia
- Compromiso social

2 Acceso Abierto

[A]n old tradition and a new technology [which had] converged to make possible an unprecedented public good [T]he willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge.

Orígenes

El movimiento del Acceso Abierto no es único, sino que responde a varias iniciativas que surgen por parte de distintos colectivos y de manera simultánea:

- 1. Los bibliotecarios y el uso abusivo de los *big deals* por parte de las editoriales científicas**
- 2. Los físicos y la velocidad de los descubrimientos científicos**

Algunos hitos importantes



1991 - Paul Ginsparg launches ARXIV



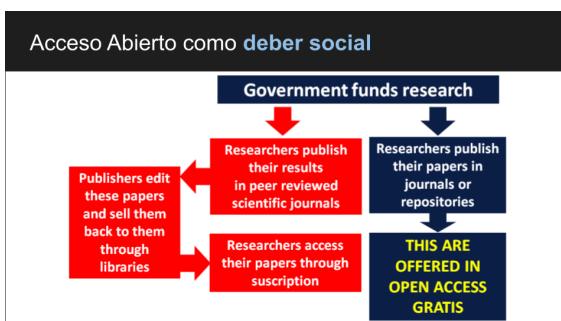
2002 - Budapest Open Access Initiative

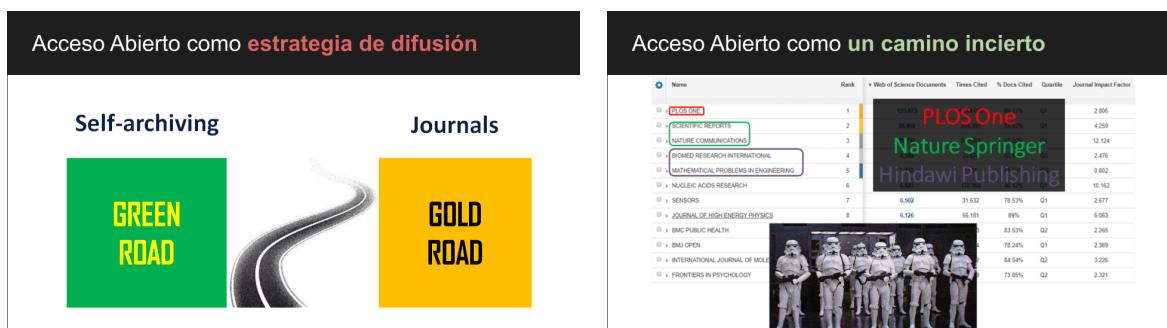


2002 - Doris Lessig develops the Creative Commons licenses

Aquí es importante destacar cómo todas estas diferentes corrientes empiezan a encontrarse y organizarse hasta encontrar un modelo de negocio que permita un equilibrio entre el negocio editorial y la libertad del investigador para hacer públicos sus resultados.

Razones para publicar en abierto





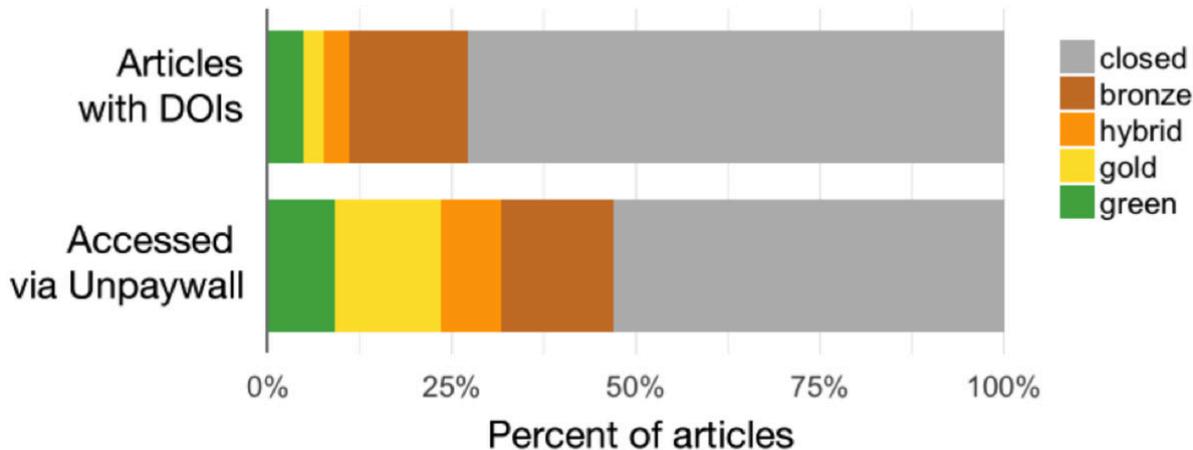
Acceso Abierto y bibliometría

Líneas de investigación abiertas:

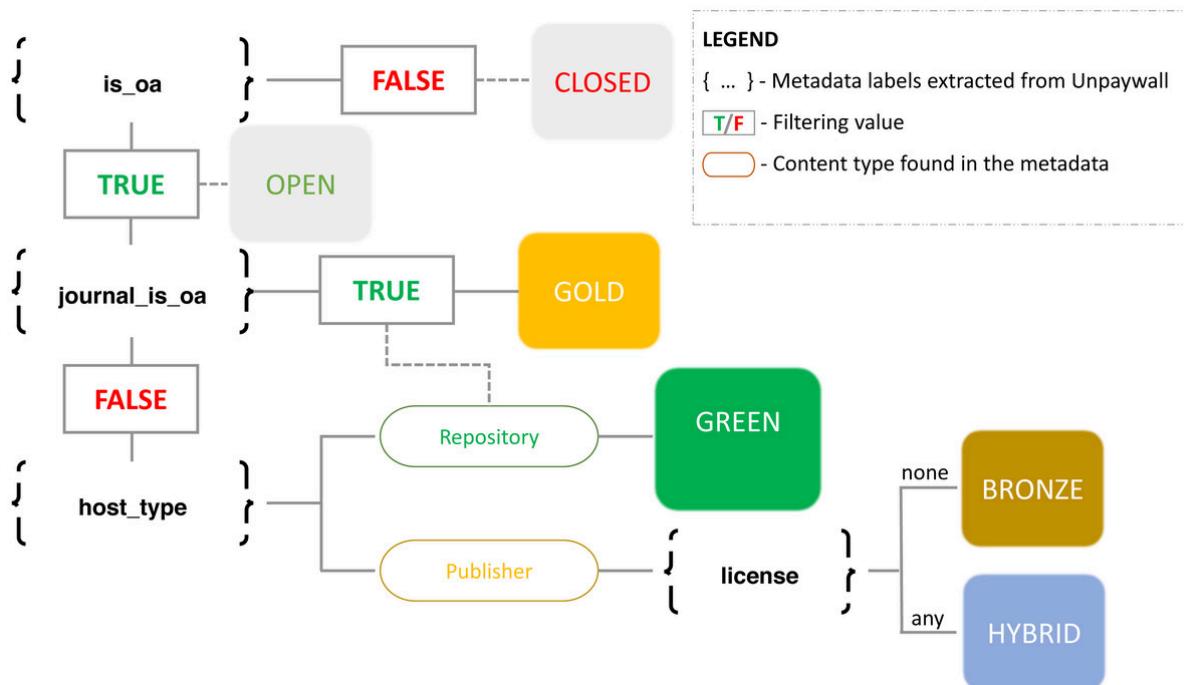
1. **Acceso e impacto:** [Harnad & Brody, 2004](#) - [Moed, 2007](#) - [Gargouri et al., 2010](#) - [Bautista-Puig et al., 2020](#)
2. **Monitorización del Acceso Abierto:** [Archambault et al., 2014](#) - [Piwowar et al., 2018](#) - [Martín-Martín et al., 2018](#) - [Robinson-Garcia et al., 2020](#)
3. **Costes y calidad de publicación:** [Beall, 2012](#) - [van Noorden, 2013](#) - [Bohannon, 2013](#)

¿Cómo definir en términos operacionales qué es acceso abierto?

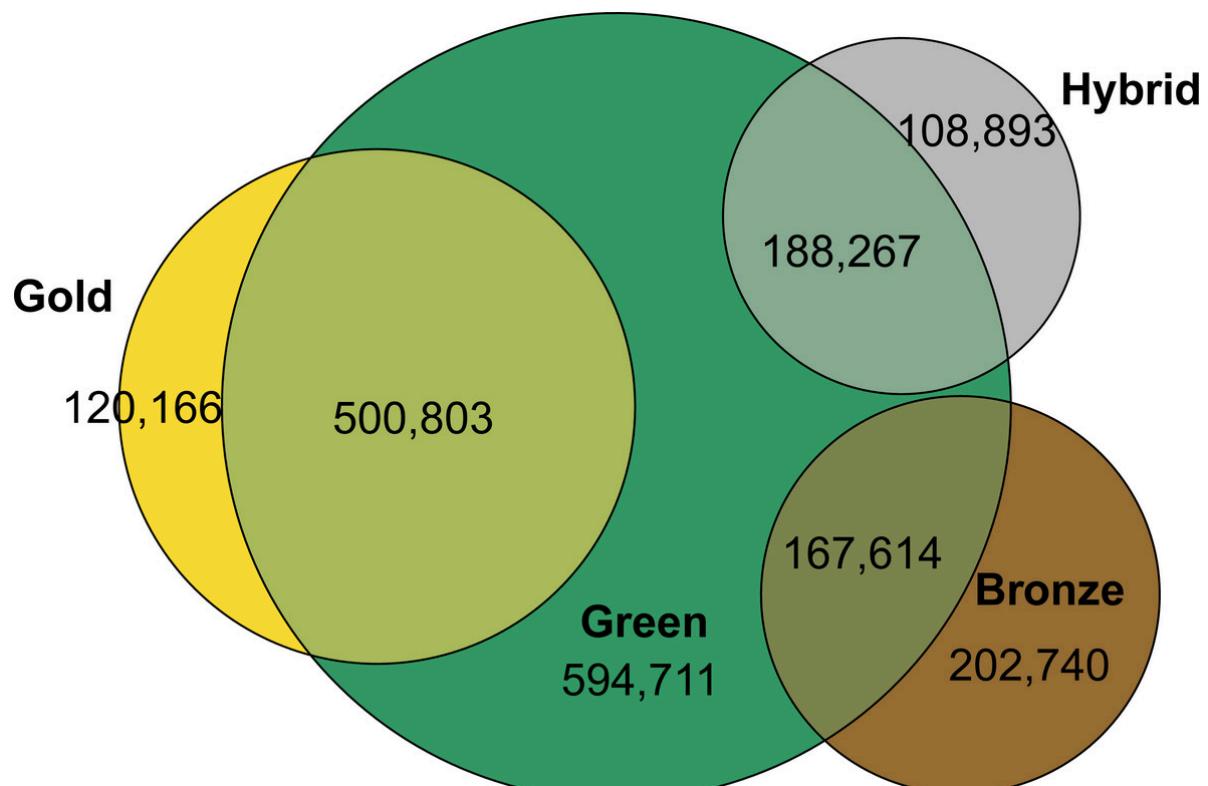
Un documento para estar en Acceso Abierto debe estar **1) libre de derechos de autor** y su acceso debe ser **2) universal** y **3) estable** en el tiempo. En principio hay 2 tipos de Acceso Abierto: **dorado** y **verde**. Pero con el tiempo la cosa se ha complicado. Ahora se habla de **bronze**, **híbrido**, **diamante** o **negro**.



Piwowar et al., 2018



Robinson-Garcia et al., 2020



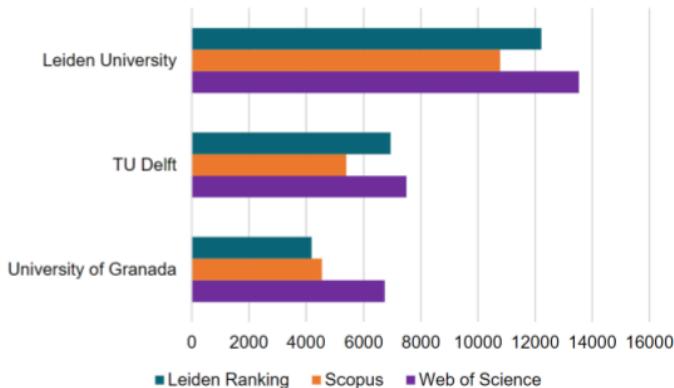
Robinson-Garcia et al., 2020

Unpaywall: el estándar actual

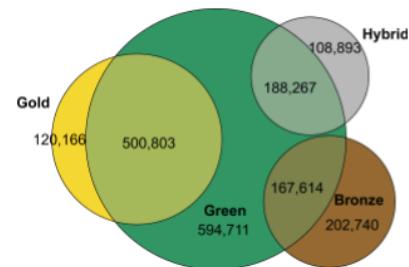
Se trata de la principal fuente que se emplea hoy día para identificar versiones en abierto de cualquier trabajo. Esto se debe a que se trata de la única herramienta que hace esto posible hoy día.

Integration with other data sources

Unpaywall has become a *de facto* standard for identifying OA literature



Different definitions + coverage may difficult interpretation of numbers



Al igual que sucede con las altmétricas, tiene una gran dependencia en DOIs.

Main features

- ★ Developed by non-profit **Our Research**
- ★ It tracks OA versions of published research with a DOI identifier
- ★ It offers information of location of OA versions and licensing
- ★ It tracks repositories, publishers and uses secondary data sources such as DOAJ or Cross

An open database of 32,990,449 free scholarly articles.

We harvest Open Access content from over 50,000 publishers and repositories, and make it easy to find, track, and use.

[CANCEL JOURNAL SUBSCRIPTIONS](#) [GET THE EXTENSION](#)

Metadata

Data Format

Overview

The [database snapshot](#), [Simple Query Tool](#), [REST API](#), and [Data Feed](#) products all return JSON-formatted data. For simplicity, that data is organized under the same schema in all cases; that schema is informally described on this page.

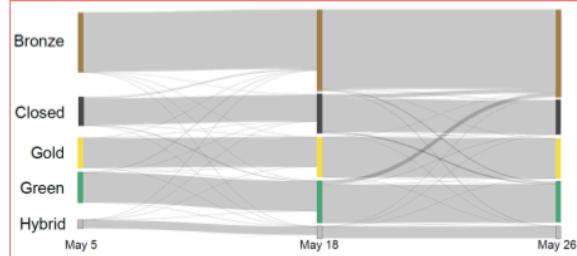
Regardless of the source, each record returned consists of one [DOI Object](#), containing resource metadata. Each DOI Object in turn contains a list of zero or more [OA Location Objects](#).

New fields may be added at any time. This won't be a problem for existing code in most cases since they will simply go unused, but you shouldn't rely on the number of fields being fixed.

Fields marked (beta) may have their behavior changed without warning. Changes to other fields will be announced on the [Unpaywall mailing list](#).



Unpaywall tracks data sources constantly, but does not necessarily keep track of changes in OA status. Although marginal, we do observe some volatility on its data.



Su funcionamiento es como el de un [motor de búsqueda](#). Es decir, no almacena información histórica, sino que rastrea la web periódicamente buscando versiones en abierto de un documento.

Aunque la mayor parte de las veces se utiliza únicamente para ver tipologías de Acceso Abierto, sus datos son muy ricos y podrían dar lugar a nuevas métricas como por ejemplo, el porcentaje de documentos depositados en el repositorio institucional que produce una universidad.

University	Country	Pubs	Green pubs	% Green pubs in Repository ^a
Bilkent Univ	Turkey	2,008	1,858	97.7–97.7
City Univ London	United Kingdom	2,569	2,131	88.3–88.6
Durham Univ	United Kingdom	7,452	6,159	84.9–85.1
Hong Kong Polytech Univ	China	9,816	7,925	5.1–96.2
London Sch Hyg	United Kingdom	7,237	5,817	76.2–76.7
Univ Strathclyde	United Kingdom	4,847	3,830	88.9–89.0
Univ St Andrews	United Kingdom	5,780	4,497	79.2–79.7
Loughborough Univ	United Kingdom	4,274	3,271	85.9–86.2
Univ Pretoria	South Africa	6,432	4,873	93.7–93.7

Data retrieval process

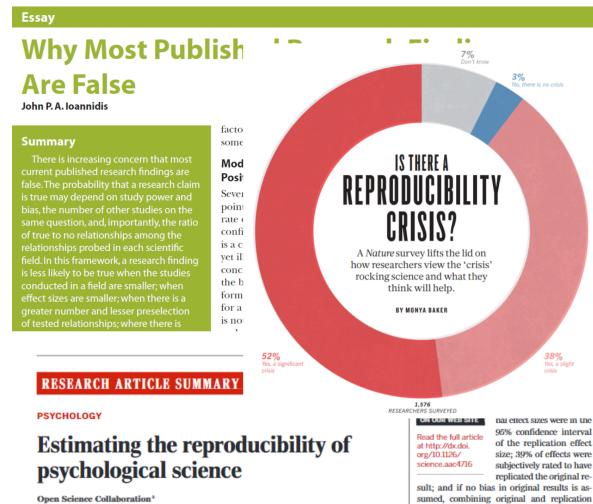
Unpaywall includes 4 options for data collection

1. Through an API. Computer skills needed! [Here a way to do it.](#)
2. Email requests. Send the list of DOIs and they will email the data
3. Through a Chrome Addon. This allows accessing to OA versions of published literature.
4. Dump file. You can actually download the whole Unpaywall dataset. Although currently it has over 100 GB!

3 Los datos de investigación

Un [polémico ensayo publicado en PLOS Biology](#) en 2005 proclamaba que la mayor parte de los resultados de investigación en el área biomédica eran irreproducibles. Esto se debía a una falta de transparencia en los métodos de los trabajos y a la toma de una serie de decisiones arbitrarias que podían poner en peligro la salud de pacientes.

Esta crítica se hizo extensible a otros campos, en especial a la [Psicología](#), como resultado del descubrimiento de varios casos de fraude científico así como falta de replicabilidad que habían pasado totalmente inadvertidos por la comunidad.



Data sharing e impacto científico

Normalmente, los investigadores son reticentes a compartir sus datos de investigación. Por lo que desde el principio se hizo un esfuerzo por buscar incentivos para que los investigadores comparten datos de investigación y herramientas metodológicas (software, código, etc.).

Heather Piwowar es un nombre con el que os tenéis que quedar, esta investigadora hizo su tesis doctoral sobre este tema y con la intención de demostrar que a mayor transparencia, habría un mayor impacto científico, aunque no excesivo.

Value all research products

A new funding policy by the US National Science Foundation represents a sea-change in how researchers are evaluated, says **Heather Piwowar**.

Fuentes de datos de investigación: Data Citation Index y DataCite

Uno de los principales problemas para monitorizar el impacto en citas de los sets de datos es la calidad de los metadatos así como los hábitos de citación de los investigadores.

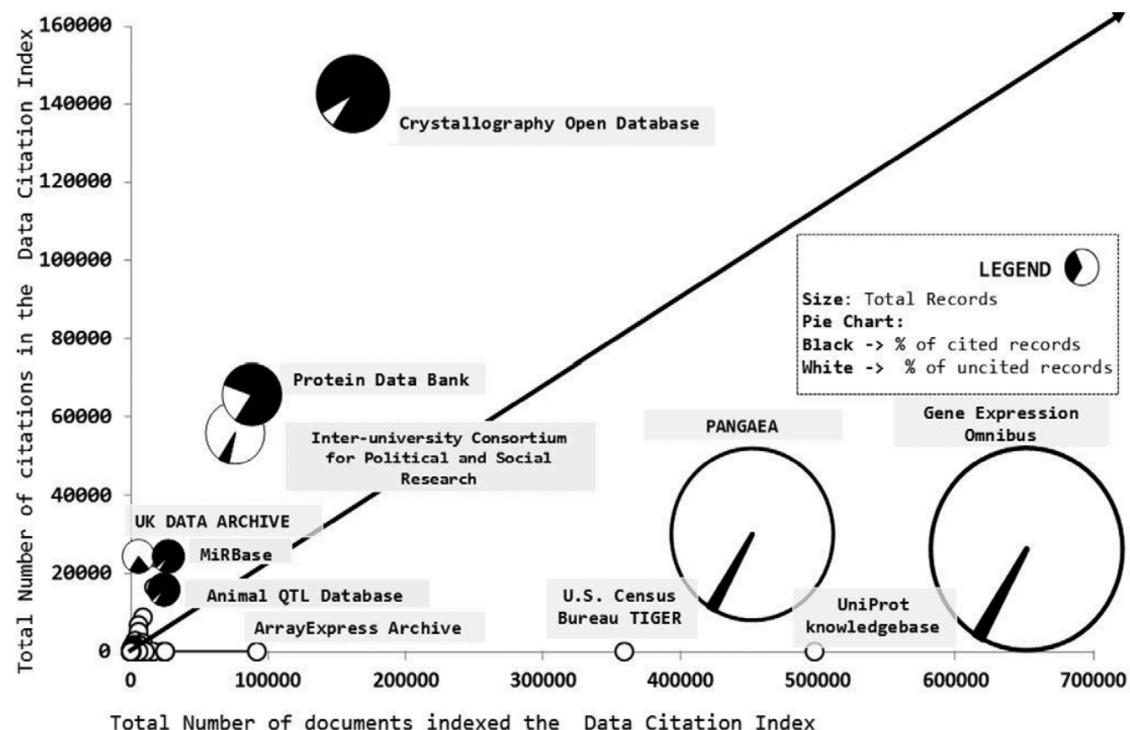
process is available in an R Notebook at

https://github.com/Wences91/social_media_communities. All methods have been automated and gathered under the R package ‘altanalysis’ (<https://github.com/Wences91/altanalysis>).

Based on this population of Twitter users, we create a matrix of relations (the data behind the network) between each member with the script *friendsnet.R*. We made the script publicly available at the GitHub repository ([Robinson-Garcia 2017](#)).

friendsnet.R first searches for the interaction community of a given user and then

En 2012, entonces Thomson Reuters (ahora Clarivate), lanzó el **Data Citation Index**. Esta base de datos, aún activa no ha tenido una gran acogida, pero fue un primer esfuerzo bastante interesante por crear índices de citas asociados a datos de investigación.



Otro esfuerzo importante es que hace **DataCite**, una plataforma encargada de dotar de DOIs a sets de datos y crear y procesar los metadatos necesarios para su compartición. Esta plataforma totalmente abierta, permite la descarga de registros para su análisis posterior. Sin embargo, tras analizarlo, se observa que aún sigue siendo muy difícil determinar el nivel de citación que reciben los sets de datos. Por otra parte, existe una gran variedad de maneras de publicar datos en abierto y citarlos, haciendo aún más complejo el tema.