Bilbao, octubre de 2021

**EF--¿?. Acceso a fuentes externas de publicaciones**

Contents

[**Introducción**](#_heading=h.30j0zll) **2**

[**Clasificación de los metadatos de una publicación**](#_heading=h.3znysh7) **4**

[Análisis de fuentes externas de publicaciones](#_heading=h.ruqdufwxpj47) **7**

[Web of Science](#_heading=h.2et92p0) 7

[ORCID](#_heading=h.tyjcwt) 9

[PubMed](#_heading=h.3dy6vkm) 10

[Semantic Scholar](#_heading=h.1t3h5sf) 11

[arXiv](#_heading=h.4d34og8) 12

[Crossref](#_heading=h.2s8eyo1) 13

[Google Scholar](#_heading=h.3rdcrjn) 14

[Open Citations](#_heading=h.ici73mdfo10u) 14

[Conclusiones sobre las fuentes analizadas](#_heading=h.26in1rg) 22

[Fuentes principales y complementarias](#_heading=h.eieuwjvpv1p8) **24**

[**PudMed**](#_heading=h.35nkun2) **24**

[**Semantic Scholar**](#_heading=h.44sinio) **24**

[**CrossRef**](#_heading=h.2jxsxqh) **25**

[Open Citations y European PMC](#_heading=h.dscp7vi7cnjw) 25

[**Compatibilidad de fuentes.**](#_heading=h.bnw9n6ji2tt4) **29**

[Análisis de la compatibilidad entre fuentes externas](#_heading=h.4i7ojhp) 29

[Análisis de la compatibilidad de la fuente principal y complementaria](#_heading=h.2xcytpi) 31

[Compatibilidad de metadatos](#_heading=h.1ci93xb) 31

[Gestión de publicaciones incompletas](#_heading=h.3as4poj) 33

[**Otros ROs**](#_heading=h.109itu5x1xab) **34**

# **Introducción**

El objetivo de este entregable es realizar un análisis inicial de otras fuentes externas de ROs de tipo publicación que puedan ser una alternativa a Scopus-Elsevier, fuente externa para la que ya se ha desarrollado un microservicio en Hércules y que posee una serie de restricciones de cuotas o número de llamadas a la misma que obligan a realizar este estudio. Para ello, se analizarán las diversas fuentes externas, sus cuotas, sus restricciones de uso, los metadatos devueltos por cada publicación y otras características determinantes en la elección de la próxima fuente externa.

La principal finalidad de este entregable es encontrar la fuente externa que más se adecue a nuestras necesidades, y por tanto nos permita, con el mínimo número de *requests (peticiones)* posible, obtener los metadatos necesarios para describir todas las publicaciones que tiene un investigador en esa fuente externa. De este modo también se tendrá en cuenta si los identificadores, tanto de autores como de artículos obtenidos, son propios de dicha red (como el *scopus\_id* de Scopus), inexistentes (por ejemplo, nombres) o identificadores generales (ORCID y DOI).

Una vez elegida la fuente externa se implementará un servicio que nos permita acceder a ella y reclamar los datos de un investigador, tal y como ya se ha hecho para la fuente externa Scopus-Elsevier.

## 

# **Clasificación de los metadatos de una publicación**

A continuación, se listan los metadatos a obtener de cada publicación. Aquellos que están coloreados no podrán ser modificados por el usuario:

* Tipo de publicación**[[1]](#footnote-1)**
* Título
* Abstract
* Idioma
* DOI
* Fecha
* Link
* Palabras clave
* Autor correspondiente
  + Nombre (lista)
  + ORCID
  + Identifier (ID en caso de no tener ORCID)
  + link
* Conjunto de autores (por cada uno):
  + Nombre (lista)
  + ORCID
  + Identifier (ID en caso de no tener ORCID)
  + link
* Página inicio
* Página fin
* Revista
  + Nombre
  + ISSN
  + ¿métrica?
* Métrica de publicación
  + Nombre de red
  + Número de citas.

Por otro lado, todos los artículos citan y referencias a otros. Dada una publicación principal como la de arriba tenemos que poder enlazarla con los artículos citados y referenciados, que serán considerados **ROs de tipo referencia o bibliografía.** Estos artículos citados y referenciados pueden ser caracterizados con los mismos metadatos que un RO de tipo publicación, pero no será necesario recuperar inicialmente todos ellos. En el siguiente listado se indican con colores los metadatos mínimos necesarios para caracterizar un RO de tipo referencia o bibliografía.

* Tipo de publicación
* Título
* Abstract
* ID
* Idioma
* DOI
* Fecha
* Link
* Autor correspondiente:
  + Nombre (lista)
  + ORCID
  + Identifier (ID en caso de no tener ORCID)
  + link
* Conjunto de autores (por cada uno):
  + Nombre (lista)
  + ORCID
  + Identifier (ID en caso de no tener ORCID)
  + link
* Página inicio
* Página fin
* Revista
  + Nombre
  + ISSN
  + ¿Métrica?
* Métrica de publicación
  + Nombre de la fuente externa
  + Número de citas

El listado de colores significa lo siguiente:

* Rojo – Metadatos necesarios e IMPRESCINDIBLES por ser identificadores estándar.
* Naranja - Metadatos importantes que nos son identificadores.
* Negro – Metadatos de tipo enlace o con información complementaria.

# **Análisis de fuentes externas de publicaciones**

En esta sección se expondrán diversas fuentes externas que contienen ROs de tipo publicación. En las siguientes subsecciones se detallarán cada una de ellas. Al final de la sección se expondrá una tabla en la que de manera resumida se expondrán los metadatos, en comparación con los deseados, que se obtienen de cada publicación, las APIs que permiten obtener estos metadatos, las ventajas y contras de estas APIs y, por último, una valoración respecto a la utilidad de dicha fuente externa. De esta manera, se podrá decidir qué nuevos microservicios de consumo de fuentes externas se deberán implementar en Hércules.

## Web of Science (WoS-Clarivate)

Esta fuente externa pertenece a Clarivate y está accesible a través de la [licencia nacional que gestiona la FECYT para las universidades españolas](https://www.recursoscientificos.fecyt.es/servicios/acceso-metadatos). En su web indican que para poder usarla hay que ponerse en [contacto con FECYT](https://www.recursoscientificos.fecyt.es/contacto) a través de un formulario en el que se indica qué tipo de servicio desea habilitar.

Una vez hecha la solicitud, el procedimiento técnico para acceder a esta fuente externa exige estar registrado, dar de alta una aplicación desde la que se accederá a las APIs, subscribirse a cada una de las APIs que se van a usar para poder recibir posteriormente por correo electrónico una API-key para poder invocar a cada API. En la sección [*How to get started*](https://developer.clarivate.com) de su web se explican los pasos a seguir.

El [portal de Clarivate](https://developer.clarivate.com/apis/wos-journal) proporciona varias APIs, además de las específicas de WoS. En concreto para obtener información de Web of Science ofecen las siguientes opciones:

* + - * [InCites Document Level Metrics API](https://developer.clarivate.com/apis/incites) – Da información sobre las citas de un artículo.
* [Web of Science API Expanded](https://developer.clarivate.com/apis/wos) - Support 'search' and 'data integration' using Web of Science data returned as JSON or XML. This API supports rich searching across the Web of Science to retrieve item-level metadata, including times cited counts and contributor addresses/affiliations. Additional operations support the ability to discover related records as well as cited references and citing items.
* [Web of Science API Lite](https://developer.clarivate.com/apis/woslite) - This API supports simple searching across the Web of Science to retrieve core item-level metadata.
* [Web of Science SUSHI API](https://developer.clarivate.com/apis/sushi-api) -
* [Web of Science Reviewer Locator API](https://developer.clarivate.com/apis/wosrl-api) - Combining the power of the Web of Science's author and reviewer databases, Web of Science Reviewer Locator is a full stack solution to find, screen, and connect with expert peer reviewers.
* [Web of Science Journals API](https://developer.clarivate.com/apis/wos-journal) - This API provides journal-level metadata and metrics for all journals in the Journal Citation Reports™ covered in the Web of Science Core Collection, including the Journal Impact Factor™ and other new metrics. Integrate journal data into your internal systems or retrieve journal indicators for bibliometrics studies.

La [licencia nacional que gestiona la FECYT para las universidades españolas](https://www.recursoscientificos.fecyt.es/servicios/acceso-metadatos) permite el acceso a la “Web of Science API Expanded”, por lo que para realizar la primera pregunta y obtener información sobre los artículos de un autor se usará la consulta por identificativo de autor AI=({ORCID\_autor}) que ofrece esta API. La funcionalidad completa de esta API se encuentra descrita en este [enlace](http://images.webofknowledge.com/WOKRS529AR7/help/WOS/hp_advanced_examples.html). El número de ítems devueltos en cada petición debe estar entre 1 y 100, por lo que para aquellos usuarios que tengan más de 100 publicaciones registradas en WoS deberá realizarse la misma petición tantas veces como sea necesario. Sin embargo, no se ha encontrado información respecto a cuotas o límite de *requests* con una determinada suscripción, aunque se seguirá indagando por si esta información está disponible en alguna sección de su web que aún no hemos encontrado.

En la tabla se encuentran diversos puntos a favor y en contra de esta fuente externa. En la tercera columna encontramos las APIS que se estiman necesarias para la obtención de los metadatos que deseamos. Cada una de estas APIs tiene un color con el que se ha coloreado en la última columna los metadatos que se obtienen con ella. Esta información está sujeta a cambios, ya que la información obtenida por las APIs no ha podido ser contrastada con una llamada.

## ORCID

Esta [fuente externa](https://orcid.org/) puede consultarse en el siguiente enlace. Para [acceder a la API](https://info.orcid.org/es/documentaci%C3%B3n/tutoriales-de-api/Api-tutorial-leer-datos-en-un-registro/#easy-faq-2606) que nos proporciona esta fuente externa se debe solicitar un token, vale con que sea de tipo *read-public* y no es necesario ser miembro para solicitarla. Es de tiempo prolongado, es decir, no caduca hasta dentro de 20 años y menciona que se pueden hacer varias llamadas con este mismo *token* pero no comentan si el número de llamadas es ilimitado o no.

Solo hay una API que permite, dado un autor, reclamar la información que se encuentra en los diferentes apartados (o todos) de su perfil. La información de cómo reclamar estos datos está en el segundo enlace de esta subsección. No hay ningún ejemplo de qué metadatos se obtienen de cada artículo, pero dado que en la página web no hay demasiados metadatos por artículo, es posible que no obtengamos la información básica necesaria que deseamos para cada artículo. Este es el problema principal de esta fuente externa. Además, no está claro si esta es la fuente externa en la que los usuarios suelan tener la información actualizada de sus publicaciones.

El resumen de la información obtenida de esta fuente externa sigue el mismo esquema descrito para la fuente externa WoS-Clarivate en la tabla de final de sección.

## PubMed

Las APIs de esta fuente externa son gratuitas y no se necesita una API-Key para acceder a las misma:

1. NCBI E-utilities --- <https://www.ncbi.nlm.nih.gov/books/NBK25501/> Po lo que tengo entendido con esta API puedes preguntar por todos los recursos de PubMed y de algún otro repositorio.
2. File validation tools, --- <https://www.ncbi.nlm.nih.gov/pmc/pub/validation/>
3. OA web service, --- <https://www.ncbi.nlm.nih.gov/pmc/tools/oa-service/>
4. OAI-PMH, --- <https://www.ncbi.nlm.nih.gov/pmc/tools/oai/> - Especifica de PubMed.
5. ID converter, --- <https://www.ncbi.nlm.nih.gov/pmc/tools/id-converter-api/>
6. FTP service, --- <https://www.ncbi.nlm.nih.gov/pmc/tools/ftp/>
7. Literature Citation Exporter. --- <https://api.ncbi.nlm.nih.gov/lit/ctxp>

Esta fuente externa es de dominio exclusivo biomédico, a diferencia de las otras fuentes externas examinadas que abarcaban varios dominios de la ciencia. El problema principal de estas APIs es el número de *requests* permitidos. En concreto de la fuente que permite obtener datos de publicaciones en PubMed se comenta lo siguiente:

*“If you are using a script that makes more than 100 requests of any kind, please run it outside of the PMC system's peak hours. Also, please make sure that your system does not make concurrent requests, even at off-peak times. Peak hours are Monday to Friday, 5:00 AM to 9:00 PM, U.S. Eastern time.”*

En otras APIs que también permiten obtener información de PubMed (como es el caso de la primera), el número de *requests* permitido por segundo es 3, en comparación con otras fuentes externas que nos permiten desde 100-300 *requests* por segundo.

## Semantic Scholar

La información sobre esta fuente externa puede consultarse en este [enlace](https://www.semanticscholar.org/). Esta fuente externa proporciona diversas APIs, aunque en muchas ocasiones se puede conseguir la misma información a través de una sola API aumentando los campos del parámetro *fields* en una misma petición o *request*. La información sobre las diversas APIs y los metadatos o campos de información que se pueden obtener con cada una de ellas se puede consultar el siguiente [enlace](https://api.semanticscholar.org/graph/v1#operation/get_graph_get_paper_references). Para información más concreta sobre la APIs puede consultarse también el siguiente [enlace](https://www.semanticscholar.org/product/api#Fetch-Author).

El mayor problema de estas APIs es que los IDs que identifican los artículos y autores devueltos son propios de *Semantics Scholar* y no aquellos que se pueden considerar como generales o estándar. Tiene una cuota de *requests*, pero dejan a disposición de cada organización o institución un formulario para rellenar en caso de que se quieran ampliar la cantidad de peticiones o *requests* por segundo que se van a realizar. Por lo demás, es gratuita y no necesita una API-key. Tampoco es posible determinar el ID de la revista en la que ha sido publicado un artículo, ya que únicamente devuelve una cadena de caracteres o *string* con el nombre de la revista. También es imposible realizar la búsqueda de los artículos de un cierto autor sin el id propio que se le otorga en esta fuente externa. Esta búsqueda tampoco puede hacerse con el ORCID de un investigador.

Al igual que con las anteriores fuentes externas, en la tabla adjunta a esta sección puede encontrarse información de las diversas APIs a utilizar, los metadatos obtenidos con ellas y los pros y contras de esta fuente externa.

## arXiv

Esta fuente externa puede encontrarse en el siguiente [enlace.](https://arxiv.org/) Ofrece APIs de acceso público, pero hay un aviso sobre el uso comercial de las mismas que es conveniente revisar, ya que habría que informar a arXiv del uso que se pretende hacer y ello puede tener implicaciones en el modelo de negocio o de explotación de las herramientas desarrolladas en Hércules:

*“Commercial projects that utilize arXiv’s public APIs or other bulk data pipelines are requested to contact arXiv at nextgen@arxiv.org, review arXiv's brand guidelines, sign a memorandum of understanding, and consider becoming an arXiv affiliate before embarking on the project. This includes any project created as a product for sale.”*.

Esta fuente externa es diferente a las examinadas anteriormente, ya que permite declarar como publicación recursos que no han sido publicados en una revista científica. Además, no tiene perfil de los usuarios, por lo que los identifica únicamente con el nombre pudiendo dar lugar a problemas de identificación.

Esta fuente externa no está resumida en la tabla debido a que se deja a los revisores de este informe la decisión de seguir considerándola como fuente externa candidata a incluir en Hércules, dado el tipo de publicaciones que se pueden encontrar en ella.

Si se decide abordar el consumo de esta fuente externa, podría considerarse la posibilidad de utilizar algún servicio desarrollado en ASIO que permitiera realizar la identificación de un usuario a partir de múltiples nombres que pueda tener en diferentes fuentes externas.

## Crossref

Esta fuente externa puede consultarse en este [enlace](https://www.crossref.org/). La documentación de las APIs proporcionadas y sus límites de cuotas, ya que tiene un número máximo de *requests* por segundo, puede encontrarse en el siguiente [enlace](https://api.crossref.org/swagger-ui/index.html). En las primeras pruebas realizadas para obtener todos los artículos de un autor a partir de su ORCID, las respuestas proporcionaban más información de la deseada, incluyendo artículos que no publicados por dicho autor. Por tanto, parece recomendable usarla simplemente para complementar la información de un cierto artículo, es decir, como fuente externa secundaria o complementaria.

En cuanto a los metadatos obtenidos de un artículo, los relativos a los autores no incluyen información que los defina de forma unívoca como investigador, como por ejemplo el ORCID u otro ID propio de la red, sino que únicamente están definidos por su nombre, pudiendo producirse errores en su identificación. Además, los metadatos relativos a las referencias o bibliografía del artículo no son suficientes ni aclaratorios.

Cabe destacar que los servidores de esta fuente externa se cayeron cuando se realizaron diferentes pruebas de acceso, por lo que es posible que los resultados obtenidos y las conclusiones a las que hemos llegado estén sesgados por esta incidencia. También se encontraron algunos valores extraños en algunos metadatos, como por ejemplo un 0 en el número de artículos referenciados en la bibliografía, lo que refuerza el comentario anterior.

**[ACTUALIZACIÓN]** Al realizar nuevas pruebas de búsqueda, con el ORCID de un usuario únicamente se han obtenido publicaciones de dicho usuario, aunque su número era mucho menor que el de las obtenidas en fuentes externas como Scopus-Elsevier o WoS-Clarivate. El análisis de los metadatos obtenidos para las publicaciones está aún en curso.

## Google Scholar

No ofrece una API para acceder a su base de datos porque Google solamente desea ofrecer esta información a través de una interfaz web para humanos. Aún así hay un desarrollo **no oficial de Google Scholar** llamado [SerpApi](https://serpapi.com/) (Google Search API) que tiene una [API (google-scholar-author-articles)](https://serpapi.com/google-scholar-author-articles) para obtener las publicaciones de un autor. El problema es que está basado en *web-scraping* por lo que la información de cada artículo se limita a lo que ofrece la interfaz web de Google Scholar y sujeta a las posibles modificaciones que haga Google en dicha interfaz. Aunque Google asigna a cada investigador un ID interno, los datos de los autores que se obtienen tras llamar a la API son sus nombres, no un ID genérico ni el identificador único propio de Google Scholar. Esta API solo devuelve 100 resultados por búsqueda, por lo que habría que realizar varias peticiones para aquellos autores con más de 100 artículos.

Para poder realizar una consulta que permita obtener las publicaciones de un autor es necesario saber el *author\_id* del investigador en Google Scholar, que se puede obtener invocando a la [API (google-scholar-profiles-api)](https://serpapi.com/google-scholar-profiles-api). Por lo tanto, surge la necesidad de filtrar los resultados con el nombre del investigador y posteriormente identificar al autor correcto y obtener el *author\_id* que le corresponde.

La [API (google-scholar-author-citation)](https://serpapi.com/google-scholar-author-citation) para saber más sobre el número de citas y la revista en la que está publicado un artículo necesita el *author\_id* de Google Scholar que devuelve la llamada inicial a la [API (google-scholar-author-articles)](https://serpapi.com/google-scholar-author-articles).

Este desarrollo no oficial exige obtener una API-key para consumir sus APIs, pero no está descrito el proceso para obtenerla, si bien hay un enlace para [registrarse en la web de SerpApi](https://serpapi.com/users/sign_up).

## Open Citations

Dado el DOI de un artículo, esta fuente externa permite obtener los DOIs de los artículos que referencia en su bibliografía o que los que lo citan a él. No aporta mucha más información, es decir, no hay muchos más metadatos. Las APIs que permiten obtener esta información se encuentran en los siguientes enlaces:

* [https://opencitations.net/index/coci/api/v1#/references/{doi](https://opencitations.net/index/coci/api/v1#/references/%7Bdoi)}
* [https://opencitations.net/index/coci/api/v1#/citations/{doi](https://opencitations.net/index/coci/api/v1#/citations/%7Bdoi)}

1. European PMC

Esta fuente externa puede encontrarse en el siguiente enlace: <https://europepmc.org/>. Esta fuente externa tiene el propósito de permitirnos obtener información sobre los artículos que citan y referencian a uno dado. El conjunto de APIS que aparecen a los consumidores puede encontrarse en el siguiente enlace:

<https://europepmc.org/RestfulWebService#!/Europe32PMC32Articles32RESTful32API/citations>

En nuestro caso lo que nos interesa son las siguientes APIs:

* <https://europepmc.org/RestfulWebService#!/Europe32PMC32Articles32RESTful32API/references>
* <https://europepmc.org/RestfulWebService#!/Europe32PMC32Articles32RESTful32API/citations>

Para poder usarlas es crucial saber el ID del artículo en esta página web, para ello hay que hacer una petición extra al siguiente url, colocando el DOI en donde corresponda. <https://www.ebi.ac.uk/europepmc/webservices/rest/search?query=DOI:10.1007/bf00197367>

Los artículos citados o referenciados aparecen con metadatos básicos entre los que se encuentra el DOI, razón por la cual esta fuente externa nos puede permitir enlazar una Ro de tipo publicación con sus ROs de tipo bibliográfica y cita. En caso de que estas peticiones listadas con anterioridad no nos den suficiente información sobre un artículo citado se deberá ejecutar la primera petición y completar los fields necesarios.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fuente Externa** | **Pros y Contras** | **APIs interesantes/necesarias** | **Metadatos** |
| Web of Science  Nota: No se ha podido probar ninguna petición, por lo que la comprobación de los metadatos recuperados son los indicados no es fiable al 100%. | [Contra] Hay que declarar específicamente qué APIs se van a usar porque cada una tiene una API-key diferente. | API para saber las publicaciones de un autor:  Web of Science API Expanded | DOI  Título  Abstract  Tipo  Fecha  Autores /autor principal -> Names  Nombre  Id  Knowledge área  Etiquetas  Número de artículos en la bibliografía  Número de citas (refs)  Palabras clave  Links  Página de inicio y fin  Revista ID  ID, (a confirmar)  nombre  Metadatos de revista (¿)  Bibliografía ->  Metadatos de cada artículo |
| [PRO] Los metadatos obtenidos son muy extensos |
| [Contra] El número de ítems devuelto en la API es limitado. | API para obtener información de las métricas de una revista:  Web of Science Journals API |
| [Contra] Son de pago. Hay que registrarse, pedir permiso y realizar una suscripción. |
| [PRO] WoS tiene muchas publicaciones y es una red muy enriquecida y muy usada. |
| Web of Science: Como organización hay que mirar si es posible/aceptable, dadas las condiciones del proyecto, solicitar estas APIs. Los metadatos obtenidos son muy completos, obteniendo IDs generales, así como una gran gama de metadatos. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **ORCID** | [Contra] Pocos metadatos de las publicaciones. | Solo tienen [una API](https://info.orcid.org/es/documentaci%C3%B3n/tutoriales-de-api/Api-tutorial-leer-datos-en-un-registro/#easy-faq-2606) que es la que se usaría para saber los artículos de un autor (la primera a realizar).  En azul se pone la información que se obtiene de esta API o al menos que está en la página web asociada a cada artículo. | Doi  Título  Abstract  Tipo  Fecha  Autores /autor principal ->  Nombre,  Id  Knowledge área  Etiquetas  Número de artículos en la bibliografía  Número de citas  Links  Página de inicio. Página fin  Revista ->  ID,  nombre,  Metadatos de revista (¿)  Bibliografía ->  Metadata de cada artículo |
| [Contra] No tiene diferentes APIs para poder completar la información obtenida. |
| [PRO] Conseguir el Token de acceso es fácil y además este dura un largo periodo de tiempo. |
| [--] Hay que considerar que no todos los usuarios declaran en esta fuente sus artículos. |
| ORCID: Devuelve pocos metadatos y los usuarios han de tenerlo al día. En general la información devuelta es muy escasa y por tanto se necesitaría extraer información adicional de otras fuentes para completar la información. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **PubMed** | [Contra] Es de un dominio de la ciencia en específico… Es de medicina no está pensado para otros campos. | Dadas las restricciones y los contras tan grandes se ha decidido no verificar qué metadatos se pueden extraer de esta fuente externa. | Doi  Título  Abstract  Tipo  Fecha  Autores /autor principal ->  Nombre  Id  Knowledge área  Etiquetas  Número de artículos en la bibliografía  Número de citas  Links  Página de inicio Página fin  Revista ->  ID,  nombre  Metadatos de revista (¿)  Bibliografía  Metadata de cada artículo |
| [Contra] Formato en el que devuelve la información es XML y no JSON. |
| [Contra] Cuota excesivamente restrictiva que no nos sirve y que además te bloquean el usuario en caso de sobrepasar el límite. |
| PubMed: No es viable dadas las imposiciones de *requests* exigidas, además de centrarse en un dominio específico y no general como otras fuentes externas examinadas.  Comentario: Igual es posible completar los datos de publicación de scopus ya que este devuelve el id de PubMed asociado a una publicación. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Semantics Scholar  ([link](https://www.semanticscholar.org/product/api#Fetch-Author)) | [--] Tiene cuota de cantidad requests por segundo, pero se puede aumentar rellenando un formulario (no sé si harán pagar o no…). | La primera API es la que daría información sobre los papers de un autor dado su id.  <https://api.semanticscholar.org/graph/v1>/author/{author\_id}/papers | Doi  Título  Abstract  Tipo  Fecha year  Autores /autor principal  Nombre  Id -En su red  Homepage  Aliases  externalIds  Knowledge area  Etiquetas  Número de artículos en la bibliografía  Número de citas  Links  Página de inicio Página fin  Revista ID 🡪 venue  ID,  nombre  Metadatos de revista (¿)  Bibliografía  ID Semantic Scholar  url  title  year  authors  número de citas |
| [PRO] No se necesita ninguna API-Key ni token para reclamar la información. |
| [Contra] No tiene el id identificativo de una revista. Por lo que esta información ha de ser completada con otras fuentes externas. | También tiene otras APIs para saber más información sobre un autor.  <https://api.semanticscholar.org/>  graph/v1/author/{author\_id}?fields=…  Esta API es necesaria para obtener el id del autor del que queremos la información (la llamada de arriba). |
| [Contra] Los IDs devueltos son en la mayoría de ellos de su propia red. Al menos devuelve un identificador. |
| [PRO] Los metadatos de la bibliográfica son igual de extensos que los de la publicación principal. |
| Opinión: el sistema de llamadas es muy fácil y cómodo. La información que devuelve no es muy extensa pero sí mínima y suficiente. La mejor opción es que esta fuente externa se use como complementaria a Scopus, por ejemplo. También sirve para realizar la primera llamada, aunque hay ciertos datos que se deberían como el DOI de una publicación o el ORCID de un usuario. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Crossref**  **(se les ha caído el servidor)** | [Contra] Tiene un límite de *requests* por segundo. Creo que se puede aumentar, pero hay que declararlo. | La Api que obtiene las publicaciones de un autor. Esta petición proporciona demasiada información incluyendo artículos ajenos a dicho autor | Doi  Título  Abstract  Tipo  Fecha  Autores /autor principal (orden)->  Nombre  Id  Knowledge área  Etiquetas (un poco raras…)  Número de artículos en la bibliografía -> dudoso porque pone 0.  Número de citas (is-referenced-by-count)  Links  Página de inicio Página fin  Revista ID  ID,  Nombre  Metadatos de revista (¿)  Bibliografía  Metadata de cada artículo |
| **[Contra] No creo que sirva para obtener los artículos de un autor en concreto.** |
| **[contra] Los identificadores de persona no son devueltos, únicamente su nombre. Hay casos en los que sí se devuelve el ORCID, pero solo en algunas personas.** |
| **[contra] Tampoco devuelve mucho de la información de la bibliografía.** | API para la obtención de metadatos asociados a un autor  [/Works/{DOI}](https://api.crossref.org/swagger-ui/index.html#/Works/get_works__doi_) |
| **[PRO] En algunos artículos de info del doi de la bibliografía.** |
| CrossRef: No es la red con más información que podamos encontrar. Es muy dudoso que te ponga que la bibliografía es 0. (Puede que esto haya sido provocado porque se ha caído el servidor). La página web funciona extremadamente mal. Es necesaria más información sobre la bibliografía e identificadores generales. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Google Scholar** | [Contra] La llamada de *autor articles* da muy poca información sobre los artículos que ha publicado un investigador. Parámetro inicial: autor\_id | Si el autor no sabe su autor\_id hay que usar *Google Scholar Profiles API*, para lo cual se busca con el nombre y se debe filtrar el resultado hasta dar con el autor deseado. | Doi  Título  Abstract  Tipo  Fecha year  Autores /autor principal (orden)->  Nombre  Id  Knowledge área  Etiquetas (un poco raras…)  Número de artículos en la bibliografía  Número de citas  Links  Página de inicio Página fin  Revista ID  ID,  Nombre  Metadatos de revista (¿)  Bibliografía  Metadatos de cada artículo |
| [Contra] No se puede obtener información sobre la revista, su DOI, bibliografía o autores. Y lo obtenido no es genérico por lo que no permite búsquedas en otras fuentes. | Google Scholar Author Articles API -- Información sobre los artículos que ha publicado un autor en concreto. |
| [Contra] Si el usuario no sabe su autor\_id entonces es difícil que con la API de profiles se obtenga este valor. | La segunda llamada con el id de la primera obtendríamos  **(**[**https://serpapi.com/google-scholar-author-citation**](https://serpapi.com/google-scholar-author-citation)**)** |
| Google Scholar: La obtención de metadatos es bastante limitada y hay que hacer uso de muchas APIs para tener la información deseada. Además, las entidades no son devueltas con IDs identificativos por lo que dificulta el intentó de completar estos metadatos con otros. | | |

## Conclusiones sobre las fuentes analizadas

Por tanto, en este entregable se han analizado diversas fuentes externas con el objetivo de posteriormente realizar un micro-servicio en esta fuente que nos permita obtener todas las publicaciones de un investigador. Se han considerado las restricciones de estas fuentes externas, tales como el número de *requests* permitidos por segundo, la facilidad de uso o los procesos necesarios para obtener las credenciales necesarias. También se ha adjuntado información de las APIs que estas fuentes externas ponen a disposición de los investigadores con el fin de que estos puedan reclamar su trabajo. Se han analizado dichas APIs con el propósito de definir cuáles de ellas son necesarias para obtener los metadatos deseados, así como se ha obtenido que combinación de APIs necesaria para definir los ROs de tipo publicación en la red de Hercules.

Este análisis se ha concluido con una tabla en la que cada fuente externa examinada ha sido valorada. En ella se exponen los puntos a favor y en contra de cada fuente, las APIs necesarias y los metadatos, en comparación con los deseados, obtenidos con cada una de ellas.

En este análisis se han obtenido los problemas más relevantes en relación con cada fuente externa:

* **WoS**: Suscripción y pago de las APIs.
* **ORCID**: metadatos insuficientes y metadatos identificativo de los autores no genéricos
* **Semantic Scholar y PubMed**: imposibilidad de buscar las publicaciones de un autor. **Estas fuentes externas deberían usarse para completar la información de una publicación.**
* **Semantic Scholar**: Metadatos identificativos propios no genéricos además es imposible conseguir el DOI de las publicaciones.
* **Crossref**: Pendiente de revisión, ya que debido a que la caída de sus servidores no ha proporcionado información fiable. Algunos usuarios son identificados con el ORCID y otros no.

Por tanto, la fuente externa más recomendable en cuestión de cantidad de metadatos proporcionados e identificación de publicaciones y autores de manera genérica es WoS. Sin embargo, las fuentes externas PubMed o Semantic Scholar podrían usarse para completar la información obtenida de otras fuentes como Scopus. Ambas fuentes externas tienen problemas. En concreto PubMed debido a la limitación en el número de peticiones y Semantic Scholar debido a la falta del metadato DOI identificativo de las publicaciones. CrossRef queda a la espera de una segunda revisión cuando sus servidores funcionen correctamente.

# **Fuentes principales y complementarias**

En este apartado se va a justificar por qué ciertas fuentes externas no pueden ser usadas como principales a la hora de realizar consultas sobre publicaciones de un autor, sino únicamente como fuentes complementarias para obtener información que permita enriquecer los metadatos de cada una de las publicaciones obtenidas a través de una consulta previa a una fuente principal.

Tal y como se ha analizado previamente, las fuentes externas que se consideran como principales serían **Scopus, WoS y ORCID**, ya que permiten realizar consultas de publicaciones de un autor a través de un identificativo estándar como ORCID ("Scopus Author ID" y "WoS ResearcherID" puede estar asociado a ORCID) y además identifican las publicaciones recuperadas con su DOI, un identificador estándar y genérico con el que poder realizar nuevas consultas en otras fuentes externas para obtener más metadatos complementarios.

En las siguientes secciones se exponen diversas fuentes externas y se comentará si se va a usar o no, así como que tipo de fuente complementaria es cada una de ellas.

## PudMed

Esta fuente externa no puede ser principal porque su API no permite la consulta de publicaciones de un autor a través de ningún tipo de identificador estándar de autor (p.ej. ORCID). No existe un identificador interno para cada investigador en esta fuente externa por lo que no permite buscar por un autor, únicamente cuando se busca por su nombre. Además, los metadatos obtenidos de cada publicación son mínimos y no genéricos, por lo que no aportan información que pueda complementar a la que ya se dispone por una reclamación previa a una fuente principal que ya ha permitido obtener los metadatos mínimos de una publicación.

Además la mayoría de los artículos que se encuentran en esta plataforma se pueden encontrar en otras fuentes externas como scopus. También tiene una forma de buscar por autor ya

no hay una manera de buscar correctamente por el autor,

## Semantic Scholar

Esta fuente externa no puede ser una fuente principal porque no permite pedir información de publicaciones de un usuario a través de su ORCID, si bien permite realizar dicha consulta a través de un “author\_id” interno que cada investigador tiene asociado en esta fuente externa. Sin embargo, cuando devuelve información de las publicaciones de un autor, esta no incluye el DOI (identificativo estándar), lo que imposibilita complementar la información de la misma con otras fuentes externas. Por lo tanto, esta fuente externa solamente podría ser complementaria.

## CrossRef

Esta fuente externa tampoco permite realizar peticiones con el ORCID de un autor. Además, tampoco se ha encontrado un identificativo interno de cada usuario en esta fuente externa. Por lo tanto, tampoco se podría considerar como principal.

## Open Citations y European PMC

Open Citations:

Esta fuente de datos tiene la gran limitación de que solo ofrece la conexión con el DOI de los artículos que cita o referencia uno dado.. Sin embargo, este metadato, como se verá en posteriores análisis, es difícil de encontrar por lo que esta fuente es primordial para conseguir esta relación entre documentos. Además al devolvernos el DOI de una artículo nos permite completar los metadatos de esta publicación con otra fuente externa.

European PMC:

Al igual que ocurre con open citations esta fuente está diseñada para obtener metadatos de los ROs de tipo referencia o cita. En este caso la búsqueda es un poco más enrevesada ya que en primer lugar se debe obtener el ID en la página de esta fuente externa. Posteriormente preguntar tanto por las citas como por las referencias de este ID.

1. **Propuesta de proceso de reclamación de publicaciones**

A continuación, se describen los pasos que se llevarán a cabo durante el proceso de reclamación de publicaciones de un determinado autor.

**[1]** La persona usuaria provee alguno de los IDs que permitirán acceder a las fuentes externas consideradas como principales:

* Scopus Author ID
* WoS ResearcherID
* ORCID

NOTA: Además, podrá proporcionar una **fecha** para filtrar la consulta y reclamar solamente las publicaciones a partir de dicha fecha.

**[2]** La persona usuaria deberá seleccionar la **fuente principal** sobre la que quiere reclamar sus publicaciones **(Scopus, WoS o ORCID). Cuestión a mirar.**

[3] Se obtendrá de la fuente principal seleccionada una lista de cada una de sus publicaciones. Recuerdo que las fuentes principales siempre devuelven el DOI de cada publicación. Por cada una:

[3.1] Si este DOI pertenece a una publicación que ya está en nuestra red:

[3.1.1] Si en nuestra red solo tenemos la información mínima de ese artículo como si fuera un **RO de tipo referencia o cita** -> Paso 3.2

[3.1.2] Si este articulo esta completamente definido en nuestra red -> Proceso de fusión (está por definir).

[3.2] Si no está en nuestra red o está como una referencia/cita.

[3.2.1] ¿La fuente principal elegida nos ha devuelto todos los metadatos de la publicación principal que queremos ?

*SCOPUS nos falta información de los autores, por lo que la respuesta es un NO.*

*WOS ¿?*

*ORCID nos devuelve muy poca información de los artículos por lo que la respuesta es NO.*

[3.2.1.1] Si. -> PASO 3.2.2

[3.2.1.2] No -> complementar con otra fuente externa una o varias.

[3.2.1.2.1] CrossRef y te devuelve las referencias.

Guardar las referencias para combinar con el paso 3.2.2

[3.2.1.2.2] **Semantic Scholar.** Esta petición se debe hacer sin pedir información de las citas y referencias ya que queremos que el DOI del artículo referenciado este si o si en nuestra red!

[3.2.2] Llegados a este punto tenemos toda la información necesaria de la publicación principal pero nos falta saber más sobre **referencias**. En caso contrario hay que hacer una excepción o pactar algo

[3.2.2.1] Open Citations:

[https://opencitations.net/index/coci/api/v1#/references/{doi](https://opencitations.net/index/coci/api/v1#/references/%7Bdoi)}

(igual se puede completar la información de citas con EuropeanPMC)

[3.2.2.1.1] Si este articulo no esta en nuestra red:

[3.2.2.1.1.1] Completar la información del artículo referenciado con **Semantic Scholar** sin pedir referencias ni citas.

[3.2.2.1.2] Si está como RO publicación. -> Paso 3.2.3

[3.2.2.1.3] Si está como RO de tipo referencia ->Paso 3.2.3

[3.2.3] Llegados a este punto tenemos toda la información necesaria de la publicación principal pero nos falta saber más sobre **citas**.

[3.2.3.1] Si se eligue Open Citations: [https://opencitations.net/index/coci/api/v1#/citations/{doi](https://opencitations.net/index/coci/api/v1#/citations/%7Bdoi)}

(igual se puede completar la información de citas con EuropeanPMC)

[3.2.3.1.1] Si este articulo no esta en nuestra red:

[3.2.3.1.1.1] Completar la información del artículo referenciado con **Semantic Scholar** sin pedir referencias ni citas.

[3.2.3.1.2] Si está como RO publicación -> Paso 4.

[3.2.3.1.3] Si está como RO de tipo referencia -> Paso 4.

**[4]** Incorporación de las publicaciones al grafo de conocimiento de Hércules.

# **Compatibilidad de fuentes.**

Estas fuentes externas serán aquellas a las que no se puede acceder directamente con el ORCID de un investigador para reclamar sus publicaciones, pero que servirán para complementar la información de una publicación obtenida desde una fuente principal **cuando no se hayan obtenido los metadatos mínimos para describirla**.

Tal y como se ha analizado previamente, las fuentes complementarias que se utilizarán serán **Semantic Scholar, PubMed y Crossref. Scopus y** **WoS** también se podrán usar como complementarias cuando la primera consulta no haya devuelto todos los metadatos mínimos de una publicación y haya que realizar una segunda consulta más concreta a través de su DOI.

## Análisis de la compatibilidad entre fuentes externas

Algunas de estas fuentes externas complementarias poseen APIs (p. ej. Web of Science Journals API) que también permiten obtener más información sobre la revista en la que ha sido publicada una publicación.

Esta invocación ha de hacerse de forma automática teniendo en cuenta diferentes compatibilidades para determinar qué fuentes complementarias se podrán usar en cada caso:

* **Compatibilidad de la fuente principal y la complementaria.** Esta compatibilidad se basa en que la fuente principal haya devuelto el metadato necesario para realizar una consulta en la fuente complementaria.
* **Compatibilidad de metadatos.** Dados los metadatos obtenidos en la fuente principal ha de elegirse la fuente complementaria que permitirá recuperar los metadatos no obtenidos previamente. Es decir, aquella que permita complementar mejor la información de la fuente principal.
* **Capacidad de las fuentes complementarias.** Este análisis es el que se ha realizado en una sección anterior y se refiere a la capacidad de las fuentes externas para tener los artículos que van a ser reclamados. De este modo se podrá valorar qué fuente complementaria usar en caso de que haya varias opciones.

A continuación, se detalla cómo se analizarán las compatibilidades mencionadas. Cabe mencionar que se trata de un **análisis preliminar** realizado a partir de la documentación oficial de las fuentes y de algunas interfaces web de prueba que estas ofrecen.

Es necesario realizar un análisis más exhaustivo basado en pruebas reales cuando se puedan realizar los registros en algunas APIs, además de contemplar si las restricciones de uso de algunas APIs son compatibles con el uso que se les quiere dar en Hércules. En concreto, está pendiente:

* El proceso de registrarse para acceder a la API de **ORCID** es complejo y hay varias opciones ([Registering a Member API Client](https://info.orcid.org/documentation/integration-guide/registering-a-member-api-client/) o [Registering a Public API client](https://info.orcid.org/documentation/integration-guide/registering-a-public-api-client/)). Estamos analizando los requisitos de cada una de las opciones y recabando la información que solicitan en su web.
* Determinar si las severas restricciones respecto al reducido número de consultas que permite **PubMed** la hacen candidata a fuente externa utilizable en Hércules.
* **WoS** aún no ha contestado a la solicitud de registro y no tenemos acceso programático vía API-key, por lo que no se ha podido probar realmente si los resultados que se obtienen coinciden con los ejemplos de su documentación y la interfaz web de la misma. Hay muchos metadatos mínimos necesarios para una publicación que debemos comprobar si son devueltos o no, como por ejemplo los relativos a la bibliografía del artículo.

## Análisis de la compatibilidad de la fuente principal y complementaria

En la siguiente tabla se muestran en las filas las fuentes externas principales y en las columnas las fuentes externas complementarias, indicando para cada una de ellas el nombre del metadato identificativo necesario para realizar búsquedas (en las principales ID de usuario y en las complementarias IDs de publicaciones).

Si la fuente principal proporciona el metadato identificativo necesario para obtener más información sobre una publicación en la fuente externa complementaria, entonces el fondo del cuadrante que une dichas fuentes será verde indicando compatibilidad. Si no, será rojo.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | CrossRef  (**DOI**) | PudMed  (**PudMedID**) | Semantic Sholar (**DOI**) | Scopus 2ª llamada **(Scopus\_ID)**  **(DOI)** | WoS revista  (**ISSN**) | Open citation  (DOI) | European pm  (DOI) |
| Scopus 1ª llamada (**ORCID**) |  |  |  | Scopus\_ID |  |  |  |
| WoS (**ORCID**) |  | ¿ |  | DOI |  |  |  |
| Orcid (**ORCID**) |  | ¿ |  | DOI | ¿ |  |  |

## Compatibilidad de metadatos

En este análisis se han examinado los metadatos que se puede obtener en las fuentes externas principales y complementarias, con el fin de ver que fuente complementaria es más adecuada para completar la información que devuelve cada fuente principal.

Para ello se han creado tres tablas, una por cada fuente principal, cuya estructura es la siguiente:

* En la primera columna aparecen todos los metadatos a obtener para una publicación.
* La segunda columna pertenece a la fuente principal. En cada celda se indica con un 1 si es posible obtener el metadato de la fila correspondiente o un 0 en caso contrario.
* Las siguientes columnas van de dos en dos. Cada par de columnas pertenece a una fuente complementaria:
  + Las celdas de la primera columna indican con un 1 si es posible obtener el metadato de la fila correspondiente de la fuente complementaria o un 0 en caso contrario.
  + En las celdas de la segunda columna se realiza una suma, por fila, de los valores obtenidos entre la fuente principal y la fuente complementaria. Si el valor es mayor que 1 (celda con fondo verde), entonces ese metadato es recuperable, ya sea con la fuente principal o con la complementaria. Obviamente, un 0 (celda con fondo rojo) indica que ese metadato no es recuperable con esa combinación de fuentes.
  + En la última fila se ofrece un recuento de cuántos metadatos se obtienen en cada fuente o en la combinación de las fuentes (casillas moradas).

Dado que estas tablas ofrecen gran cantidad de información, se documentan en ficheros Excel aparte.

## Gestión de publicaciones incompletas

En caso de que con la combinación de fuente principal y fuente complementaria no se llegue a cubrir la información necesaria para registrar una publicación en el grafo de conocimiento de Hércules, se analizará la posibilidad de utilizar varias fuentes complementarias a la fuente principal. Por ejemplo, si la información de la bibliografía de una publicación no está completa, se podrían invocar a microservicios de fuentes externas complementarias a través de los DOIs de los artículos presentes en la bibliografía.

Posteriormente, habrá que determinar cómo gestionar las colisiones de metadatos que se puedan producir. La gestión se realizará de manera automática o manual, mostrando una interfaz al usuario para el tratamiento de dichos metadatos

# Otros ROs

Existen a su vez otros ROs (Dataset, Presentación, Gráficos, Docs, Enlaces, Video, Poster, Lesson) cuyos elementos son en definitiva objetos identificados con un DOI. Que a su vez serán referenciados o citados por otros ROs. Dado que su identificativo es este y que también están referenciados y citados por otros el algoritmo diseñado para la obtención de estos ROs es en definitiva el mismo. Los principales cambios son las fuentes principales que ahora serán:

* + Zenodo: Permite preguntar por el orcid de un autor
  + Figshare (Dataset, Presentación, Gráficos, Docs) -> ID
  + OpenAire: Permite preguntar con el orcid de un investigador.
  + Slideshare (Presentación) (Ejemplo de presentación) -> ID

Además estos ROs no será necesario completarlos con otras fuentes externas, sino que la información obtenida en el primer momento por estas fuentes es la que los describe completamente. .

**Bibliografía**

[1] «Research Object Crate». [En línea]. Disponible en: https://www.researchobject.org/. [Accedido: 3-marzo-2021].

[2] «Figshare». [En línea]. Disponible en: https://figshare.com. [Accedido: 3-marzo-2021].

[3] «Share». [En línea]. Disponible en: https://www.share-research.org. [Accedido: 3-marzo-2021].

[4] «GitHub». [En línea]. Disponible en: https://github.com. [Accedido: 3-marzo-2021].

[5] «Research Object Crate». [En línea]. Disponible en: https://www.researchobject.org/. [Accedido: 3-marzo-2021].

[6] «CiTO, the Citation Typing Ontology». [En línea]. Disponible en: https://sparontologies.github.io/cito/current/cito.html. [Accedido: 3-marzo-2021].

[7] «Web Annotation Data Model». [En línea]. Disponible en: https://www.w3.org/TR/annotation-model. [Accedido: 3-marzo-2021].

[8] «Dublin Core». [En línea]. Disponible en: https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/ [Accedido: 3-marzo-2021].

referencias (bib) de los documentos. otra vuelta a las apis…

<https://www.crossref.org/documentation/cited-by/retrieve-citations/>

<https://developer.clarivate.com/help/oauth2_authorization_code>

<https://developer.clarivate.com/help/oauth2_implicit_grant>

Caso de uso de Hércules <https://dev.elsevier.com/ir_cris_vivo.html>

Examples of ScienceDirect and Scopus search using CORS:

ScienceDirect Search

Scopus Search

APIs

Scopus Search API

ScienceDirect Search API

Scopus Abstract Retrieval API

ScienceDirect Article Retrieval

Scopus Abstract citation count API

ScienceDirect Entitlement API

SciVal API

ScienceDirect Article Hosting Permission API

Revisar las APIs de Scopus-Elsevier <https://dev.elsevier.com/api_docs.html>

1. Este metadato no tiene por qué ser una clasificación ni exhaustiva ni extensa de tipos de publicaciones. De momento únicamente se consideran ConferencePaper, AcademicArticle y Publication (para todo lo demás). [↑](#footnote-ref-1)