OpenAlex: breve guía de consulta a través de *R*

OpenAlex: a brief guide to querying through *R*

Ángel Borrego

Como citar este artículo:

Borrego, Ángel (2024). "OpenAlex: breve guía de consulta a través de *R* [*OpenAlex*: a brief guide to querying through *R*]". *Infonomy*, 2(1), e24011. *https://doi.org/10.3145/infonomy.24.011*

Ángel Borrego
https://orcid.org/0000-0002-6462-3966
https://www.directorioexit.info/ficha575
Universitat de Barcelona
Facultat d'Informació i Mitjans Audiovisuals
Centre de Recerca en Informació, Comunicació i Cultura (CRICC)
Melcior de Palau, 140
08014 Barcelona, España
borrego@ub.edu

Resumen

OpenAlex es un índice, gratuito y abierto, de entidades interconectadas en el ecosistema de la información científica: trabajos, autores, revistas, instituciones, temas, etc. Esta guía describe sus principales funcionalidades y ofrece ejemplos de la consulta de su API a través del paquete *openalexR* desarrollado con el lenguaje de programación *R*.

Palabras clave

Bases de datos; Buscadores; Comunicación científica; Información científica; *OpenAlex*; *openalexR*.

Abstract

OpenAlex is a free and open index of interconnected entities in the scholarly information ecosystem: papers, authors, journals, institutions, topics, etc. This guide describes its main functionalities and provides examples of querying its API through the openalexR package developed using the *R* programming language.

Keywords

Databases; Search engines; Scholarly communication; Scientific information; *OpenAlex*; openalexR.

1. Qué es *OpenAlex*

En mayo de 2021, Microsoft anunció la retirada a final de año de su buscador *Microsoft Academic*[1]. Poco después, *Our Research*[2], una organización sin ánimo de lucro dedicada

a la implementación de servicios de información abiertos en el ámbito científico, informaba de la puesta en marcha de *OpenAlex*[3] (**Priem**; **Piwowar**; **Orr**, 2022).

Sus creadores definen *OpenAlex* —un nombre alusivo a la biblioteca de Alejandría— como un catálogo abierto y gratuito de documentos científicos, investigadores, revistas e instituciones, así como de las conexiones que se establecen entre esas entidades. El lanzamiento del producto se hacía efectivo los primeros días de 2022[4]. Una noticia en *Nature* describía *OpenAlex* como "an ambitious free index of more than 200 million scientific documents that catalogues publication sources, author information and research topics" (**Chawla**, 2022).

Si bien *OpenAlex* parecía plantearse inicialmente como un sustituto de Microsoft Academic —de hecho, éste y *Crossref* son dos de las principales fuentes de datos que lo alimentan—, lo cierto es que progresivamente ha integrado información procedente de servicios como *ORCID* (*Open Researcher and Contributor ID*), *ROR* (*Research Organization Registry*), *DOAJ* (*Directory of Open Access Journals*) o Unpaywall (una base de datos de documentos en acceso abierto que probablemente es el producto más conocido de *OurResearch*).

En sus primeros dos años de vida, el producto se ha consolidado como la fuente para la elaboración del *CWTS Leiden Ranking Open Edition*[5] e instituciones como la *Sorbonne Université*[6] o el *Centre national de la recherche scientifique* (*CNRS*)[7] han anunciado su adopción como fuente de datos alternativa a *Scopus* o *Web of Science*.

2. Estructura de *OpenAlex*

En *OpenAlex* se describen diferentes tipos de entidades que establecen relaciones entre ellas. Por ejemplo:

- Trabajos (*works*): artículos, libros, *datasets*, etc. que citan otros trabajos.
- Autores (*authors*): personas que han creado esos trabajos.
- Medios (*venues*): revistas y repositorios que albergan los trabajos.
- Instituciones (*institutions*): organizaciones a las que están afiliados los trabajos a través de sus autores.
- Temas (*topics*): temas que describen los trabajos.
- Editoriales (*publishers*): organizaciones que distribuyen los trabajos.
- Financiadores (*funders*): entidades que financian la investigación.

3. Cómo consultar *OpenAlex*

Existen tres maneras de consultar *OpenAlex*:

- El sitio web: *OpenAlex* dispone de una interfaz de usuario, aunque sensiblemente más limitada y con menos funciones que *Scopus* o *Web of Science* (**Codina**, 2024). Carece de un formulario de búsqueda avanzada, la página de resultados no muestra resúmenes, apenas dispone de filtros, etc.
- El *dataset*: es posible descargar la base de datos completa, si bien es una opción compleja y desaconsejada por los responsables del producto.
- La API (*application programming interface*): se trata de un software intermediario que permite la consulta de la base de datos desde nuestro ordenador. El acceso es gratuito y no requiere autenticación.

La consulta de la API se puede llevar a cabo directamente desde el navegador. A continuación, se muestran tres ejemplos de consultas correspondientes a tres tipos de entidades. Simplemente hay que introducir en el navegador la URL que se indica. Las figuras muestran un extracto del resultado a la consulta obtenido en pantalla. Si bien es posible emplear cualquier navegador para hacer la consulta, *Firefox* ofrece una salida ligeramente más atractiva que la de otros navegadores.

3.1 Información bibliográfica del trabajo con el DOI 10.3145/epi.2020.mar.03

URL: *https://api.openalex.org/works/doi:10.3145/epi.2020.mar.03*

Figura 1. Extracto de la información bibliográfica del trabajo con el DOI 10.3145/epi.2020.mar.03

3.2 Información de la autora con el ORCID 0000-0002-9056-8251

URL: *https://api.openalex.org/authors/orcid:0000-0002-9056-8251*

Figura 2. Extracto de la información de la autora con el ORCID 0000-0002-9056-8251

3.3 Información de la revista con el ISSN 2990-2290

URL: *https://api.openalex.org/venues/issn:2990-2290*

Figura 3. Extracto de la información de la revista con el ISSN 2990-2290

En los ejemplos anteriores hemos buscado información sobre una única entidad (un trabajo, una autora y una revista), pero también es posible recuperar un conjunto de entidades aplicando filtros:

Listado de trabajos publicados en 2024
URL: *https://api.openalex.org/works?filter=publication_year:2024*

Número de trabajos publicados por cada institución:
URL: *https://api.openalex.org/works?group_by=institutions.id*

Trabajos disponibles en acceso abierto
URL: *https://api.openalex.org/works?filter=is_oa:true*

Listado de autores apellidados aristarain
URL: *https://api.openalex.org/authors?filter=display_name.search:aristarain*

Listado de revistas disponibles en acceso abierto
URL: *https://api.openalex.org/sources?filter=is_oa:true*

4. Consulta de la API de *OpenAlex* a través de R

R es un lenguaje de programación orientado al análisis estadístico. Es un lenguaje muy popular ya que permite cargar diferentes paquetes o bibliotecas (*libraries*) desarrollados por la comunidad de usuarios con funcionalidades de cálculo y de representación gráfica. A pesar de que la aparición de *OpenAlex* es reciente, ya existen diversos paquetes que facilitan la consulta de su API desde *R*. Para ilustrar su uso utilizaremos el paquete *openalexR*, desarrollado por **Massimo Aria** (2023).

Las consultas a cualquier API desde *R* pueden hacerse de forma similar a los ejemplos que hemos visto en el apartado anterior. No obstante, es posible automatizar este proceso para realizar múltiples consultas. Por ejemplo, se puede generar un bucle que realice tantas consultas como DOIs queramos consultar. Para facilitar aún más el trabajo, si no tenemos conocimientos de programación, podemos recurrir a un paquete que incluya una función para realizar estas consultas sucesivas de la API. Esto es lo que hace *openalexR*.

El primer paso es instalar el paquete y abrirlo. Para instalarlo podemos seguir alguna de las opciones descritas en *https://docs.ropensci.org/openalexR*

Vamos a replicar con *openalexR* las tres primeras consultas descritas en el apartado anterior. En primer lugar, creamos un objeto en *R*, al que denominamos 'data.1', en el que guardamos la consulta correspondiente al DOI 10.3145/epi.2020.mar.03.

```
data.1 <- openalexR::oa_fetch(
        entity = "works",
        doi = "10.3145/epi.2020.mar.03")
```

La figura 4 muestra los primeros campos del registro.

```
## $ id                          <chr> "https://openalex.org/W3012538122"
## $ display_name                <chr> "Mapa de visibilidad y posicionamiento en ~
## $ author                      <list> [<data.frame[4 x 11]>]
## $ ab                          <chr> "The visibility and the search engine posi~
## $ publication_date            <chr> "2020-03-14"
## $ so                          <chr> "Profesional De La Informacion"
## $ so_id                       <chr> "https://openalex.org/S180455664"
## $ host_organization           <chr> "Ediciones Profesionales de la Informacio~
## $ issn_l                      <chr> "1386-6710"
## $ url                         <chr> "https://doi.org/10.3145/epi.2020.mar.03"
```

Figura 4. Algunos campos del trabajo con el DOI 10.3145/epi.2020.mar.03

En vez de un único DOI podemos consultar varios simultáneamente (tres en el siguiente ejemplo).

```
data.2 <- openalexR::oa_fetch(
        entity = "works",
        doi = c("10.3145/epi.2020.mar.03",
                "10.3145/thinkepi.2018.06",
                "10.3145/infonomy.24.002"))
```

En el siguiente extracto de código, indicamos que la entidad que deseamos consultar es una autora y ejecutamos la búsqueda por el identificador ORCID.

```
data.3 <- openalexR::oa_fetch(
        entity = "authors",
        identifier = "orcid:0000-0002-9056-8251")
```

La figura 5 muestra algunos campos del registro de la autora, como su nombre, número de trabajos en la base de datos, número de citas que han recibido estos trabajos, afiliación y áreas de especialización.

id	display_name	orcid	works_count	cited_by_count	affiliation_display_name	top_concepts
A5070018631	Carol Tenopir	0000-0002-9056-8251	680	6940	University of Tennessee at Knoxville	World Wide Web, Library science, Law

Figura 5. Algunos campos del registro de la autora con el ORCID 0000-0002-9056-8251

En los siguientes ejemplos vamos a construir consultas para recuperar un listado de entidades. El primer ejemplo nos permite recuperar los trabajos que incluyen en el título la expresión “information literacy” y que han sido publicados en enero de 2024.

```
data.5 <- openalexR::oa_fetch(entity = "works",
 title.search = "information literacy",
 from_publication_date = "2024-01-01",
 to_publication_date = "2024-01-31")
```

El siguiente código recupera los trabajos que incluyen la expresión “information literacy” en el título y han sido citados más de 100 veces. Obviamente, sería posible combinar la consulta anterior y ésta en una única búsqueda incluyendo todos los parámetros.

```
data.6 <- openalexR::oa_fetch(entity = "works",
 title.search = "information literacy",
 cited_by_count = ">100")
```

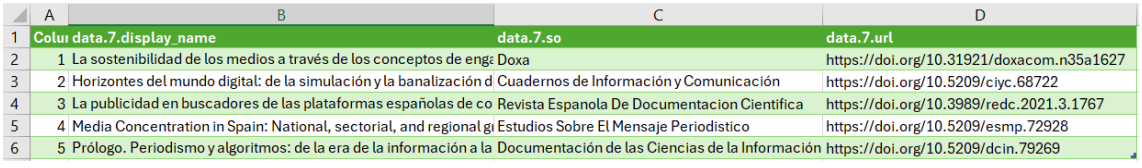
Vamos a obtener ahora la información bibliográfica de los documentos que citan el trabajo con el DOI 10.3145/epi.2020.mar.03. El identificador de *OpenAlex* que introducimos como parámetro en la búsqueda lo hemos obtenido al hacer la primera consulta (ver la primera línea de la figura 1).

```
data.7 <- openalexR::oa_fetch(entity = "works",
 cites = "W3012538122")
```

Por último, extraemos la información contenida en tres campos (título, revista y DOI) de los 5 documentos citantes y la exportamos a un fichero de texto que abrimos con *Microsoft Excel* (figura 6).

```
data.8 <- data.frame(data.7$display_name,
 data.7$so,
 data.7$url)

write.table(data.8,
 file = "info8",
 sep = ",",
 fileEncoding = 'utf-8')
```



	A	B	C	D
1	Colu	data.7.display_name	data.7.so	data.7.url
2	1	La sostenibilidad de los medios a través de los conceptos de enga	Doxa	https://doi.org/10.31921/doxacom.n35a1627
3	2	Horizontes del mundo digital: de la simulación y la banalización d	Cuadernos de Información y Comunicación	https://doi.org/10.5209/ciyc.68722
4	3	La publicidad en buscadores de las plataformas españolas de co	Revista Espanola De Documentacion Cientifica	https://doi.org/10.3989/redc.2021.3.1767
5	4	Media Concentration in Spain: National, sectorial, and regional gi	Estudios Sobre El Mensaje Periodistico	https://doi.org/10.5209/esmp.72928
6	5	Prólogo. Periodismo y algoritmos: de la era de la información a la	Documentación de las Ciencias de la Información	https://doi.org/10.5209/dcin.79269

Figura 6. Exportación a *Microsoft Excel* de los trabajos que citan el documento con el DOI 10.1007/s11192-008-2143-3

5. Disponibilidad del código

El código *R* para reproducir las consultas de este texto está disponible en *https://rpubs.com/angelborrego/openalex*

Notas

1. *https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach*
2. *https://ourresearch.org*
3. *https://openalex.org*
4. *https://blog.ourresearch.org/openalex-launch*
5. *https://open.leidenranking.com*
6. *https://www.sorbonne-universite.fr/en/news/sorbonne-university-unsubscribes-web-science*
7. *https://www.cnrs.fr/en/cnrsinfo/cnrs-has-unsubscribed-scopus-publications-database*
8. *https://github.com/ropensci/openalexR*

6. Referencias

Aria, Massimo (2023). "A brief introduction to openalexR".
https://ropensci.github.io/openalexR/articles/A_Brief_Introduction_to_openalexR.html

Chawla, Dalmeet-Singh (2022). "Massive open index of scholarly papers launches". *Nature*.
https://www.doi.org/10.1038/d41586-022-00138-y

Codina, Lluís (2024). "OpenAlex: ¿una alternativa a Scopus y Web of Science?"
https://www.lluiscodina.com/openalex-scopus

Priem, Jason; **Piwowar, Heather**; **Orr, Richard** (2022). "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts". *26th International Conference on Science, Technology and Innovation Indicators*.
https://doi.org/10.48550/arXiv.2205.01833