# ¿Qué infraestructuras abiertas?

## Which open infrastructures?

### Isidro F. Aguillo

**Aguillo, Isidro F.** (2023). "¿Qué infraestructuras abiertas?". *Anuario ThinkEPI*, v. 17, e17a21.
*https://doi.org/10.3145/thinkepi.2023.e17a21*

Publicado en *IweTel* el 16 de junio de 2023

**Isidro F. Aguillo**
*https://orcid.org/0000-0001-8927-4873*
*https://www.directorioexit.info/ficha67*

*Consejo Superior de Investigaciones Científicas*
*Instituto de Políticas y Bienes Públicos*
*Laboratorio de Cibermetría*
Albasanz, 26-28
28037 Madrid, España
*isidro.aguillo@csic.es*

**Resumen:** Ante las peticiones de reforma de los procesos de evaluación científica en el marco del desarrollo de la llamada *Open Science*, se propone la creación de una o varias plataformas abiertas específicamente diseñadas para tal fin. Se realiza una breve evaluación crítica de algunos de los servicios actualmente utilizados o disponibles y se concluye que, aunque proporcionan contenidos importantes que pueden ser reutilizados, por si solos no son adecuados para una evaluación rigurosa, fidedigna y transparente. Algunas características que dicha nueva herramienta debe cumplir son descritas y comentada su viabilidad.

**Palabras clave:** *Open Science*; Infraestructuras abiertas; Evaluación de la investigación; Bases de datos académicas; Métricas.

**Abstract:** There is a strong movement for radical reform of the scientific evaluation processes in accordance with the principles of the open science initiative. To support the new system, one or several open platforms specifically designed for evaluation purposes is badly needed. A brief critical evaluation of some of the services currently used or available is carried out, and it is concluded that, although they provide important content that can be reused, the services alone are not enough nor completely suitable for rigorous, reliable and transparent evaluation. A proposal for a new tool is introduced with a brief analysis of its feasibility, along with a description of the needs and characteristics it must comply with.

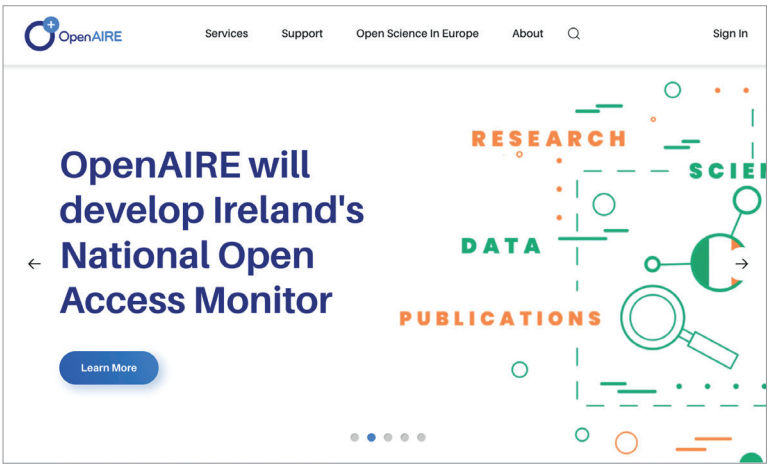
**Keywords:** Open Science; Open infrastructures; Research evaluation; Academic databases; Metrics.

## 1.Introducción

Al amparo de la implementación de políticas de Open Science[1] se están creando, reformando o definiendo una amplia serie de infraestructuras que apoyen su implantación. En Europa ello incluye entre otros:

- la puesta en marcha de la *European Open Science Cloud* (*EOSC*), la nube donde depositar y explotar contenidos;
  *https://eosc-portal.eu*
- el apoyo al gran meta-repositorio *OpenAIRE*;
  *https://www.openaire.eu*
- el desarrollo del *Open Research Europe*, una alternativa piloto a la publicación en revistas.
  *https://open-research-europe.ec.europa.eu*

En esa agenda, una importante iniciativa es la reforma de los procesos de evaluación de la investigación, cuyo diseño final influirá en el diseño u características de los actuales sistemas de información científica y su futura evolución. Un importante documento en este sentido es el Acuerdo de la *Coalition for Advancing Research Assessment* (*CoARA*, 2023) que describe algunos principios y compromisos para tal reforma.




*https://www.openaire.eu*

Este trabajo analiza los productos y servicios que podrían ser candidatos a convertirse en la herramienta de evaluación para la *Open Science*. Posiblemente ninguno de ellos individualmente cubra las necesidades específicas que garanticen los principios de transparencia, inclusión, equidad y responsabilidad, pero conociendo sus características y mecanismos podría servir de base al diseño de una o varias fuentes de datos abiertos con control documental específicamente diseñadas para satisfacer las necesidades de los futuros procesos de evaluación.

## 2. Algunos candidatos

La oferta actual de herramientas utilizadas en evaluación incluye servicios que requieren registro tales como *Mendeley*, *ResearchGate* o *Academia.edu* que se ignoran en este análisis por razones obvias. Otra limitación adicional es que, por cuestiones prácticas, esta propuesta se refiere a un posible sistema estatal de información científica para cubrir las necesidades españolas, aunque la necesidad de usar estándares abiertos sin duda permitiría futuras integraciones con otras iniciativas nacionales o supranacionales (Europa, Latinoamérica). El modelo de dicho sistema y su posible funcionamiento ya ha sido descrito previamente (**Aguillo**, 2022a), pero sin entrar en detalle técnico de la herramienta.

Obviamente habría que empezar con los proveedores comerciales, que constituyen el núcleo de la actual generación de herramientas de evaluación que se pretende superar. Se trataría de *Incites* (*Clarivate*) y *Scival* (*Elsevier*), productos muy sofisticados que presentan numerosos problemas. Es caro el acceso a la información y las condiciones de utilización no son tampoco las adecuadas, incluyendo el uso de indicadores propietarios. Quizás más grave es su cobertura limitada, excluyente y sesgada que, a pesar de su mejor control documental, es su mayor limitación. En todo caso cabría duplicar sus formatos de explotación de datos e incluso considerar como fuentes de ciertas métricas, vía API, las bases de *WoS*/*Scopus*.

Un ejemplo concreto de tal uso es el llamado *Ranking de "Stanford"* que usa un indicador compuesto con datos extraídos de la base de datos *Scopus*.
*https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/4*

Con fines evaluativos es interesante porque incluye datos en bruto reutilizables, añade multidimensionalidad por el uso de un indicador compuesto, proporciona una gran cobertura de investigadores (unas 170.000 entradas sobre un total de 8-9 millones de autores) e intenta categorizarlos por disciplinas. Sin embargo, son frecuentes los errores de filiación y disciplina y no provee de un mecanismo de corrección propio.

Más sofisticados en el uso de esos datos serían los llamados rankings de universidades, que combinan datos de distintas fuentes con información bibliométrica, por ejemplo, de *WoS* (*ARWU*, Leiden), *Scopus* (QS, *THE*, *Scimago*) o *Google Scholar* (*Webometrics*). Desafortunadamente la información no bibliométrica no está disponible en abierto y la bibliométrica aparece a menudo notablemente elaborada. En todo caso la cobertura institucional limita su uso, aunque habría datos de interés siempre de forma complementaria: Por ejemplo, posiciones a diferentes niveles de agrupación geográfica, disciplinar o temática (ODS por *THE* y *QS*, medioambiental por *GreenMetric*), información de penetración de acceso abierto, colaboración internacional o intersectorial o reparto por género, datos que proporciona Leiden, el único que suministra un fichero *Excel* con datos en bruto. Tampoco habría que descartar la información proporcionada por las propias universidades y que compila *U-Multirank*.

Fuera del duopolio habría que citar 2 productos basados en nuevas fuentes:

- *OpenAlex* se construyó con los datos del cancelado MS Academic y ha ido realizando un titánico esfuerzo de corrección y normalización de su base de datos.
  *https://openalex.org*

  En el ínterin ha ofrecido un API generoso que incluye todos sus contenidos, lo que ha sido aprovechado, respetando la licencia original C0, por Research.com para producir una serie de rankings por países y disciplinas. Sin embargo, son datos que ni se controlan, ni se actualizan con frecuencia y con cobertura limitada: Solo hay 3.500 españoles en dichas listas.
  *https://research.com*
- Un caso similar en principio, pero muy diferente, es el AD *Scientific Index* que explota datos de los perfiles de *Google Scholar*.
  *https://www.adscientificindex.com*

  La licencia de *GS* no permite su explotación y de hecho no existe API a tal efecto, aunque siempre han sido flexibles con el uso de *web scrapping* con fines de investigación. Este servicio turco va mucho más allá, no sólo ha creado un servicio con más de 3 millones de perfiles de GS, sino que abiertamente lo explota comercialmente, algo flagrantemente ilegal. El interfaz es muy visual y agradable e incluso añade una clasificación temática, aunque de los más de 34.100 perfiles marcados para España, sólo 16.700 (49%) están etiquetados con disciplinas. El control documental también es muy mejorable.

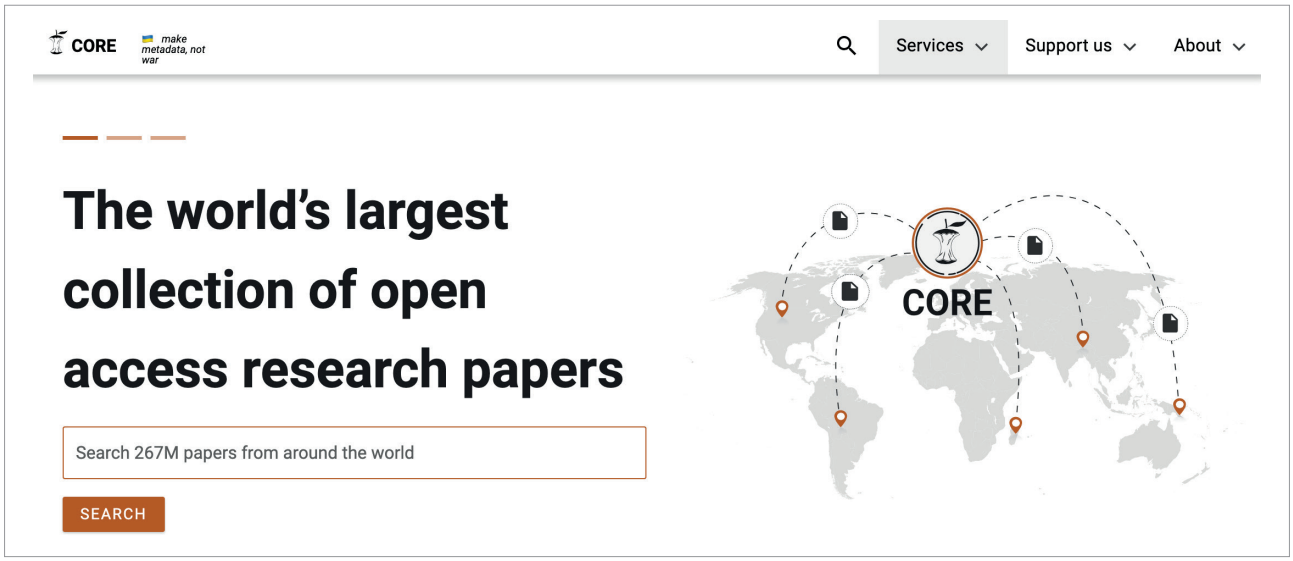
Otras fuentes con datos bibliométricos explotables, no necesariamente abiertos, son

- *Dimensions*;
  *https://app.dimensions.ai*
- *Lens*;
  *https://www.lens.org*
- *Semantic Scholar*;
  *https://www.semanticscholar.org*
- *Scilit*;
  *https://app.scilit.net*

todos ellos con un número mucho mayor de registros que *WoS* o *Scopus*.

Un grupo de candidatos con un diseño diferente serían los meta-repositorios. En la actualidad hay tres grandes servicios universales con entre 200 y más de 300 millones de registros:

- *Core*;
  *https://core.ac.uk*
- *OpenAIRE*;
  *https://explore.openaire.eu/*
- *Base*.
  *https://www.base-search.net*




*https://core.ac.uk*

A pesar de los esfuerzos de sus gestores, hay un gran número de ítems duplicados, muchos de los metadatos no están normalizados y el control documental es muy mejorable. Más importante aún es la ausencia total de métricas, aunque como en otros casos la disponibilidad de la iniciativa *OpenCitations* podría solventar dicha limitación.
*https://opencitations.net*

Es cierto que *OpenAIRE* está trabajando en la mejora de la calidad de contenidos y la interconexión de diferentes formatos, incluidos datos abiertos, pero es un proceso laborioso que en todo caso no lo cualificaría como instrumento de evaluación integral.

Finalmente hablaríamos de los CRIS o portales de investigación, o mejor dicho de la federación de CRIS, por ejemplo, a nivel estatal. En España hay distintos proveedores de software CRIS con implantaciones en varias universidades y centros de investigación.
*https://dspacecris.eurocris.org/cris/explore/dris*

Desde un punto de vista métrico son muy descriptivos y fundamentalmente explotan datos del duopolio. Dada la variedad actual parece improbable que se pudiera disponer de un CRIS nacional por fusión. Sin embargo, dada su posición como fuente con datos propios y el número de instalaciones con que los que ya cuenta, sería Dialnet (17) una probable candidata ese servicio.
*https://fundaciondialnet.unirioja.es/servicios/dialnet-cris*

La gran ventaja estaría en el control de calidad que se implementa a nivel local. Sin embargo, echamos a faltar una mucho más amplia colección de indicadores, incluyendo valores relativos y una herramienta de explotación más allá de la visualización amigable.

## 3. Una propuesta

La revisión anterior no pretendía ser exhaustiva, pero algunos patrones son claros: Los proveedores comerciales siguen siendo una fuente de datos elaborados de gran importancia que, con la correspondiente contrapartida económica, pueden ofrecer datos e indicadores de calidad para ser tenidos en cuenta como parte, ya no más como todo, en el proceso de evaluación. Los servicios públicos no cumplen necesariamente todas las opciones "open" y casi nunca son FAIR, pero además no garantizan un amplio control documental, ofrecen un limitado número de variables y no consta herramienta generadora de informes personalizados para evaluación.

Tres características echamos a faltar en esos servicios

- Una gran cantidad de variables. La oferta, no necesariamente simultanea ni completa en todos los valores, de un gran número de indicadores cuantitativos o normalizados de forma cuantitativa, quizá no menos de 200 entre absolutos y relativos. Todo ello apoyado por identificadores y metadatos lo más estándar posible sensu FAIR.
- Una gran cobertura de la producción y los productores. Por ponerlo en contexto hablaríamos de entre 200 y 250 mil autores y varios millones de documentos y otros ítems para el caso de España.
- Un sistema interactivo de auto-cálculo a varios niveles, con selección de umbrales, la posibilidad de filtrar según múltiples criterios y mecanismos de explotación y corrección supervisada. Como se ha comentado (**Aguillo**, 2022b) utilizando una bibliometría más sofisticada.

Respecto al primer requerimiento cabe asumir que los actuales desarrollos en minería de datos, el uso amplio de múltiples APIs y la contratación, en su caso, de aquellas fuentes comerciales podrían ser suficientes para su desarrollo. No obstante, de forma adicional para garantizar la calidad de los contenidos sería oportuno mantener un equipo de indizadores humanos y ofrecer de forma transparente (y quizá anonimizada para cumplir leyes de protección de datos personales) un sistema público de revisión y corrección de errores por los involucrados.

La base de datos supra-bibliográficos puede construirse por acreción de varias fuentes, algunas de las cuales han sido comentadas anteriormente. Una federación de CRIS parece una solución a medio plazo según la evolución de implementaciones en nuestro país. La liberación de la base de CVNs, el volcado de contenidos de *CrossRef*, *OpenAlex* u otras fuentes requiere invertir recursos en de-duplicación y normalización, pero no parece inviable con la inversión adecuada.
*https://www.crossref.org*

La tercera parte es la más simple técnicamente: Ante una convocatoria de evaluación concreta definida por un órgano competente, este plantea los requerimientos específicos en ese caso concreto de los filtros, umbrales y cálculos a realizar y el programa devuelve datos individuales o combinados actualizados a dicho momento.

Pero también la más abierta a polémica. En muchos casos, aunque se describan las condiciones y circunstancias que guían la obtención de las métricas finales, el evaluado podría sentir que se trata de

un sistema de tipo *black-box*. Por ello habría que hacer especial hincapié en la reproducibilidad de los resultados. Un sistema interactivo abierto plantea algunos retos, pero sería la solución óptima.

## 4.Consideraciones finales

Teniendo en cuenta que los procesos de evaluación suelen tener un fuerte carácter nacional, la viabilidad de esta propuesta habría que contrastarla en ese marco concreto, en nuestro caso España. Por ejemplo, aun teniendo en cuenta la diversidad de software CRIS implantados en nuestro país, una federación virtual podría ser posible. Alternativamente *Fecyt* dispone de la gran colección de CVNs de investigadores españoles.
*https://cvn.fecyt.es/datos-cvn/investigadores-usuarios*

Aunque supondría un gran trabajo, se podría derivar una plataforma abierta sobre la que construir un sistema como el descrito. Sin embargo, la explotación debería corresponder más bien a agencias de tipo Aneca. Es posible que una redistribución de recursos permitiera este sistema con costes extra limitados.

Por último, garantizar la calidad de la información requiere contar con instituciones, intermediarios (bibliotecarios, documentalistas) y los propios evaluados que deberían recibir formación específica, incluyendo aspectos éticos y de responsabilidad individual y colectiva.

## 5. Notas

1. Utilizo el término en inglés de Open Science, en vez de su traducción al castellano, porque el término "ciencia abierta" ha sido utilizada por distintos autores en sentidos no coincidentes con la propuesta original de la Unión Europea.

## 6.Referencias

**Aguillo, Isidro F.** (2022a). "Mejores métricas para una mejor evaluación". *Revista de la Sociedad Española de Bioquímica y Biología Molecular*, art. 781.
*https://revista.sebbm.es/articulo.php?id=781&url=mejores-metricas-para-una-mejor-evaluacion*

**Aguillo, Isidro F.** (2022b). "Bibliometría sofisticada" . *Anuario ThinkEPI*, n. 16.
*https://doi.org/10.3145/thinkepi.2022.e16e28*

*CoARA* (2022). Agreement on Reform of Research Assessment.
*https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf*

⬡ ⬡ ⬡

## Otras iniciativas

**Paloma Martín-Arraiza**

Me parece interesante la reflexión. Me gustaría añadir un comentario y dos iniciativas que no se han nombrado pero que contribuyen a esa abertura y FAIRificación de datos en el contexto bibliometrico.

Por un lado, *EOSC* no es una nube para depositar y exportar datos. El nombre, quizá, nunca fue acertado, ya que en el imaginario colectivo una nube es un *Drive* o un *AWS*. *EOSC* nunca ha pretendido ser eso, sino una federación de servicios de datos con estándares FAIR.

El siguiente artículo del año 2019 proporciona un buen contexto al respecto:

**Budroni, Paolo**; **Claude-Burgelman, Jean**; **Schouppe, Michel** (2019). "Architectures of knowledge: The European Open Science Cloud". *ABI technik*, v. 39, n. 2, pp. 130-141.
*https://doi.org/10.1515/abitech-2019-2006*

También puede ser de interés la presentación de Ignacio Blanquer, miembro del Comité Director de *EOSC Association*, en la inauguración de las *Jornadas Técnicas de RedIRIS* en Zaragoza.
*https://tv.rediris.es/es/jjtt2023/video/64817f74d2fafa003227cecb*

Luego en cuanto a las iniciativas, considero interesantes las siguientes

- *Open Citation Index* y el correspondiente identificador persistente (aún emergente) *OCI*
  *https://opencitations.net/index*
- *Open Global Citation Corpus* de *DataCite* con apoyo de *Wellcome Trust* y la Chan *Zuckerberg Initiative.*
  *https://blog.datacite.org/data-citation-corpus-announcement-2023*

**Paloma Marín-Arraiza**
*p.arraiza@orcid.org*