# Mejores métricas para una mejor evaluación

**Isidro F. Aguillo**
Responsable del Laboratorio de Cibermetría del Instituto de Políticas y Bienes Públicos del Consejo Superior de Investigaciones Científicas (CSIC).

## UNA POSTURA PERSONAL

Antes de empezar me gustaría establecer mi posición personal en dos importantes aspectos relacionados con la evaluación mediante métricas:

**1.**Rechazo el uso generalizado y poco informado de la bibliometría fácil o de salón. Es completamente inaceptable describir fenómenos tan diversos y complejos como la generación de conocimiento científico, su distribución y la descripción de su impacto recurriendo a uno o dos indicadores de una o dos fuentes de datos. Promuevo siempre el uso de múltiples fuentes y gran cantidad de variable s que profesionales expertos conviertan en indicadores significativos.

**2.**Es incorrecto utilizar las métricas en la búsqueda de la excelencia, sea como sea tanto su definición como su medida, cuando la prioridad en España es descubrir y eliminar la mediocridad. Mediocridad derivada de prácticas, no solo moral y éticamente cuestionables como el plagio, la auto-citas y redes de citantes, sino también abiertamente delictivas como el nepotismo y la endogamia, tipificables como prevaricación. Y también aquella mediocridad derivada del que, habiendo conseguido puesto fijo, no vuelve a priorizar la labor investigadora.

## ABANDONAR LAS MÉTRICAS DE REVISTAS

En el mundo se publican alrededor de 60.000 revistas con algún tipo de revisión por pares, a lo que había que sumar una ingente cantidad de monografías y un número no menor de otros formatos utilizados en la comunicación científica. En muchos procesos de evaluación que utilizan trabajos publicadas en las revistas fuente del llamado primer cuartil (Q1) eso significa que solo se evalúan los contenidos de unos pocos miles de revistas. Empíricamente se puede demostrar que un número no despreciable de dichos artículos no son citados y que estadísticamente es muy probable que se ignoren trabajos altamente citados en revistas no fuente. El uso de métricas de revistas no tiene sentido en este momento cuando ya se puede acceder directamente a las citas individuales de cada artículo. Más aún cuando además muchas de esas métricas tienen serios problemas:

a) Las fuentes *WoS* y *Scopus* incluyen solo una fracción del total de las publicaciones, un grupo con importantes sesgos de cobertura y falta de diversidad. Las ciencias sociales y las humanidades, las ciencias de la computación, las ingenierías y las ciencias "locales" están fuertemente infrarrepresentadas.

b) El factor de impacto (FI) solo recoge citas realizadas en un año lo que apenas representa una fracción no representativa del total de citas que finalmente recibirá un trabajo. Y esto es válido para todas las disciplinas, no solo para las más lentas en los procesos de cita.

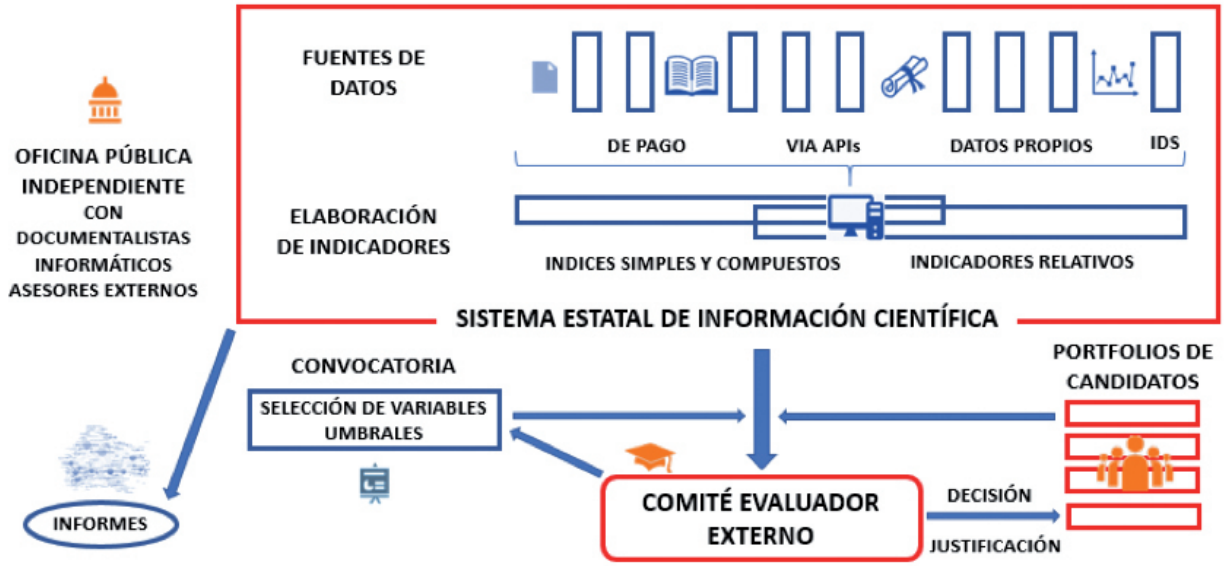
c) El FI es una medida central (media) de una distribución muy sesgada de citas, lo que la hace matemáticamente desaconsejable. Dado que el número de trabajos altamente citados es muy bajo, la mayoría (70-80%) de los artículos publicados en una revista reciben menos citas que las esperadas por su factor de impacto.

d) El FI obtenido de dividir números enteros se publica con 3 cifras decimales, una falsa precisión incluso en revistas muy citadas. Esos decimales se usan para construir cuartiles donde las diferencias entre revistas de diferentes grupos no son significativas y por tanto las agrupaciones son incorrectas e injustas.

En la actualidad existen, al menos, seis bases de datos bibliométricas con mayor cobertura que *WoS/Scopus* (*Dimensions, Google Scholar, CrosRef, Microsoft Academic, Lens, 1findr*), que proporcionan métricas de artículos más completas que las del duopolio. Sin embargo, la mayoría de ellas no alcanzan el nivel de sofisticación y control de calidad de *WoS* o *Scopus*, por lo que la recomendación es el uso informado de todas ellas conociendo sus ventajas y limitaciones respectivas.

## MÉTRICAS RESPONSABLES

Cuando hablamos de uso informado o responsable nos referimos a que las variables deben contextualizarse. ›››




››› Eso significa que es preferible no usar por defecto valores absolutos sino indicadores relativos (independientes del tamaño) y que deben considerarse factores tales como la disciplina o subdisciplina, las ventanas temporales, los ámbitos geográficos o la tipología institucional. A nivel de perfiles individuales debe calcularse la edad académica (años/lustros desde la primera publicación, excluyendo periodos inhábiles como puestos de gestión, crianza de hijos, enfermedades de larga duración), tenerse en cuenta la posición o puesto de trabajo, y otras circunstancias personales tales como el género. Por ejemplo, en el caso del frecuentemente criticado *Índice h*, tener en cuenta estos factores a la hora de calcular indicadores (la llamada familia de indicadores h) disminuye los problemas de su aplicación. Los indicadores relativos tienen también problemas si no consideramos umbrales y escenarios como lo demuestra el caso del llamado indicador de excelencia (número de artículos en el 10% más citado) que debe darse primariamente como valor absoluto y no como porcentaje si no se fijan valores mínimos de producción y se usa conteo completo.

Eso nos lleva a otros aspectos que requieren soluciones más complejas. Eliminar las auto-citas resulta injusto en la medida que muestran una línea de investigación continuada en el tiempo, pero un exceso de las mismas debe corregirse para evitar abusos. Ello requiere fijar porcentajes aceptables que, obviamente, deben ser distintos para investigadores junior (más estrictos) que los seniors y que también va a depender de las disciplinas. Aún más conflictivo es el problema de la autoría, mal resuelto en la bibliometría tradicional, tanto si se utiliza conteo completo como si se usa el fraccionado. Algunas soluciones pueden ser injustas (primar autores "principales", autor de contacto, primer o último autor), de difícil implementación (descripción del rol de cada autor) o generar rechazo (excluir la hiper-autoría, aquellos trabajos con cientos de firmantes).

Para cerrar el apartado de las citas hay que señalar que ya existen herramientas capaces de discriminar el sentido de las citas, identificando aquellas que son negativas. Esto tendrá importantes consecuencias para la bibliometría en el futuro, aunque dichas herramientas todavía no están perfeccionadas.

## LAS MÉTRICAS ALTERNATIVAS NO SON UNA SOLUCIÓN

Presentadas como alternativas a las medidas tradicionales basadas en citas, las altmétricas están lejos de ser la solución que algunos defienden. En primer lugar, no existen como un conjunto homogéneo e interrelacionado, ya que sus fuentes son muy diversas y diferentes entre sí y sus medidas cuantifican cosas distintas. Algunas como *Mendeley* o *ResearchGate* tienen componentes bibliométricos, pero mezclados con otros de interacción social fácilmente manipulables y de difícil interpretación. Twitter ofrece grandes conjuntos de datos, pero su cobertura es muy sesgada, y aunque útil para identificar comunidades, sus métricas plantean retos importantes en su aplicación a procesos de evaluación. Algunos autores afirman que se puede evaluar el impacto social a través de las menciones, estructura y valoraciones de *Facebook*, *Linkedin* y otras redes similares rusas o chinas, pero se trata de universos muy abiertos, sin metadatos controlados que permitan un análisis fino y que, en realidad, miden más interacciones que verdadero impacto. Quizá las herramientas más interesantes sean las ligadas a

depósitos, bien sea de datos (*Zenodo*), *software* (*Github*), presentaciones (*Slideshare*), videos (*YouTube*) o información (*Wikipedia*), donde la naturaleza más concreta de los contenidos y su estructura permiten derivar indicadores significativos.

Personalmente me gusta distinguir la "usometría" de otras altmétricas, en la medida que el consumo de información de cualquier sede web ofrece un enorme caudal de información, fácilmente organizable y con gran cantidad de variables de interés. Desafortunadamente la medida de visitas, visitantes o descargas están fuertemente distorsionadas por el uso de robots que hace inviable su uso en muchos casos, excepto para medir la evolución temporal, y siempre con mucha precaución.

Señalar que los productos basados en indicadores compuestos, como *Altmetric.com* no son recomendables. Las fuentes utilizadas son muy heterogéneas, priman cierto tipo de resultados (noticias de interés periodístico) de ciertas disciplinas (fundamentalmente biomédicas) y no hay una justificación clara de los pesos otorgados.

## CONSTRUIR INFRAESTRUCTURAS, UNA SOLUCIÓN DEFINITIVA

Quedaría ahora introducir mi "visión" de cómo deberían organizarse los procesos de evaluación de la ciencia, por ejemplo, en España. La propuesta se estructura en tres niveles:

1.El Sistema Estatal de Información Científica. Un órgano profesional público independiente crearía y mantendría actualizado un sistema transparente de información, detallando las actividades relacionadas con la esfuerzo y resultados de la investigación de los distintos componentes del sistema español de I+D. Se podría imaginar como una gran base de datos de los investigadores, metadatos descriptivos (edad, género, puesto, titulación, especialidad, institución) y una amplia serie de variables cuantitativas y cualitativas, de diferentes fuentes. El sistema debe permitir derivar mediante filtros y cálculos, valores representativos por grupos, umbrales según convenga e indicadores relativos diseñados exprofeso.

Las variables (docenas o más) deben guiarse por los objetivos de la *Open Science*. Es decir, además de la capacidad de producir y publicar ciencia en canales formales (revistas, libros, actas, tesis) e informales (web, redes sociales, informes), deben incluirse datos sobre el nivel de apertura de los métodos, datos y herramientas hacia los pares, el compromiso en la generación de contenidos para terceros no pares (profesores, médicos, ingenieros, otros profesionales), el esfuerzo en involucrar otros actores sociales y ciudadanos tanto en la generación (ciencia ciudadana) como en el consumo (divulgación) de resultados científicos y, por supuesto, el impacto tecnológico y económico en relación a la inversión en recursos humanos y presupuestarios realizada. Es imprescindible añadir con carácter prioritario el cumplimiento de principios éticos y morales (I+D responsable).

2.El portafolio personal. Hay que abandonar los actuales modelos de CV personales estandarizados. Frente a listados exhaustivos de actividades y resultados, deben primarse documentos personalizados donde el autor utilice sus propias palabras (narración) para describir su formación, conocimientos y principales hitos de su actividad hasta el momento. La información proporcionada debe organizarse sobre tres ejes principales donde el autor identifique (3 ó 5) resultados relevantes elegidos y justificados por él mismo: Formación, Experiencia e Impacto.

3.Los comités de evaluación. Formados por investigadores especialistas, se encargan seleccionar los criterios, variables y umbrales específicos para cada convocatoria. Las decisiones se toman contrastando los datos del sistema extraídos con los portafolios de los candidatos, justificando de forma transparente la resolución adoptada.

Este protocolo se puede adaptar más allá de casos individuales para grupos, proyectos, instituciones o regiones. Un sistema nacional permanentemente actualizado puede generar informes periódicamente a varios niveles de forma prácticamente automática. La disponibilidad de un sistema como el descrito reduciría notablemente la carga burocrática que soportan los investigadores ante cualquier proceso de evaluación, ofrece flexibilidad a los evaluadores para extraer información contextualizada en cada momento y para cada necesidad y garantiza la necesaria transparencia de dichos procesos.

El uso de múltiples fuentes y de una amplia batería de indicadores libera al sistema de costosas y, a menudo, sesgadas dependencias de monopolios extranjeros, empujando al desarrollo de fuentes propias para incorporar la medida de los impactos más locales (económicos, sociales) y aumentando la diversidad de las aportaciones susceptibles de reconocimiento. Reduciendo el peso de las métricas de revistas se puede luchar contra la inflación de los costes de publicación (APCs). Otras prioridades políticas (Plan S, Declaración de Helsinki sobre multilingüismo, políticas de igualdad, persecución de malas prácticas) pueden implementarse en la herramienta.

## PARA LEER MÁS


- DORA. Declaración de San Francisco sobre la evaluación de la investigación. 2013. https://sfdora.org/read/read-the-declaration-espanol/
- D Hicks, P Wouters, L Waltman, S de Rijcke, I Rafols The Leiden Manifesto for research metrics. *Nature*, 520 (2015): 429-431. https://leidenmanifesto.org
- D Moher, L Bouter, S Kleinert, P Glasziou, MH Sham, V Barbour et al. The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLoS Biology* 18 (2020): e3000737. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000737
- J Wilsdon, L. Allen, et al. The Metric Tide Report of the Independent Review of the Role of Metrics in Research Assessment and Management. HEFCE, London, 2015. https://responsiblemetrics.org/wp-content/uploads/2019/02/2015_metrictide.pdf