G.6. *Google Scholar*: no es oro todo lo que reluce

Por **Isidro F. Aguillo**

11 enero 2011

Aguillo, Isidro F. "*Google Scholar*: no es oro todo lo que reluce".
Anuario ThinkEPI, 2011, v. 5, pp. 211-215.

Resumen: *Luego de una corta perspectiva histórica de las bases de datos bibliográficas con citas, que permiten realizar estudios bibliométricos, se comentan las características de Google Scholar como posible base de datos para la misma utilización. Si bien Scholar es gratuito y más exhaustivo que WoS y Scopus, no se aconseja su uso para realizar análisis bibliométricos con fines de evaluación de personas e instituciones.*

Palabras clave: *Google Scholar, Google Académico, WoS, Web of science, Scopus, Bases de datos bibliográficas, Citas, Bibliometría.*

Title: Google Scholar*: all that glitters is not gold*

Abstract: *After a short historical perspective of bibliographic databases with citations, which allow bibliometric studies, the features of Google Scholar as a possible database for the same purpose are discussed. While Scholar is free and more comprehensive than WoS and Scopus, its use is not recommended for bibliometric analysis, especially for the evaluation of individuals and institutions.*

Keywords: *Google Scholar, WoS, Web of science, Scopus, Bibliographic databases, Appointments, Bibliometrics.*

Introducción

EL FACTOR LIMITANTE en los estudios de la actividad científica, especialmente los que utilizan técnicas cuantitativas, ha sido la disponibilidad de bases de datos.

La bibliometría de las últimas décadas no hubiera sido posible sin las bases de *ISI/Thomson* (*citation indexes*)[1]. La explosión de la patentometría coincide con el acceso en abierto de los servicios web de las organizaciones de patentes europeas, estadounidenses y japonesas y, en fin, la cibermetría existe en buena medida por las bondades de los motores de búsqueda comerciales.

En muchos casos dichas bases de datos no habían sido diseñadas específicamente para la actividad bibliométrica y fue necesario (y todavía lo es) realizar un importante esfuerzo de selección, limpieza, organización y normalización de los resultados, antes de comenzar cualquier análisis.

Los costes eran enormes (acceso vía *Dialog*[2], adquisición de versiones en cd-rom) y lo siguen siendo (licencias nacionales *WoK*[3] y *Scopus*), y, además de ciertas limitaciones legales, estaban las de carácter técnico. Éstas son relevantes para entender la tipología y profundidad de los trabajos bibliométricos realizados en los 80 y los 90. Era difícil exportar grandes cantidades de registros, ciertos campos tenían múltiples valores difíciles de segregar (autores, direcciones, citas), había que repasar errores y normalizar entradas, era complejo hacer correspondencias entre autores y sus direcciones cuando varios tenían la misma afiliación institucional.

La imposibilidad práctica de corresponder referencias con artículos generalizó el uso de las citas "esperadas" (el *infame* factor de impacto, por el que se supone a cada artículo particular el valor de la revista), en vez de utilizar las "observadas" o reales de cada uno. Otras consecuencias fueron el desprecio hacia los recuentos fraccionados de los cada día más frecuentes trabajos multiautorados o el insólito filtrado temático por categorías disciplinares de revistas o por selección de palabras clave (¡en bases de datos sin auténtica indización!). Todo ello motivado por las limitaciones de contenido y estructura de las bases de datos, pero también por un sistema de gestión intencionadamente capado que impedía una adecuada automatización de ciertos procesos.

La consecuencia directa es que el usuario final del trabajo del bibliómetra (otros colegas, fundamentalmente aquéllos objeto de análisis, y los gestores de instituciones y de políticas científicas)

apenas se reconoce en los resultados, que pueden pecar tanto de excesiva sencillez (plenos de errores) como de inaguantable profundidad (tablas densísimas, sin ninguna utilidad práctica).

"La dificultad de indizar la llamada internet invisible motivó la elaboración de un producto que no dependiera de los robots automáticos"

Hubo y sigue habiendo (cada vez menos, eso sí) trabajos mediocres, pero quizá la principal carencia es la ausencia de escenarios generales, con históricos de datos correctamente organizados y que evitara la continua reinvención de la rueda a la que nos tiene acostumbrada esta disciplina en nuestro país (aunque en este caso la culpa es compartida por la inaudita ausencia de un manual de calidad, actualizado en castellano, del conjunto de las disciplinas cuantitativas).

Obviamente esta nota no es la primera que llama la atención sobre el cuidado extremo que se ha de tener tanto a la hora de seleccionar las fuentes bibliográficas como en el diseño de la extracción y utilización de los datos correspondientes. Y es posible que vuelva a caer en saco roto.

Google Scholar

Para los afortunados que trabajan en instituciones que se pueden permitir el *indecente* dispendio de tener contratadas las dos grandes bases de datos de citas (*Web of science* y *Scopus*), la labor bibliométrica se hizo un poco más compleja con la aparición de éste segundo, el nuevo producto de *Elsevier*. No sólo las bases de datos eran diferentes (*Scopus* es ligeramente mayor y con menor sesgo anglosajón), sino que las herramientas de consulta y extracción y los indicadores (externos en el caso de *Scopus*) eran también distintos.

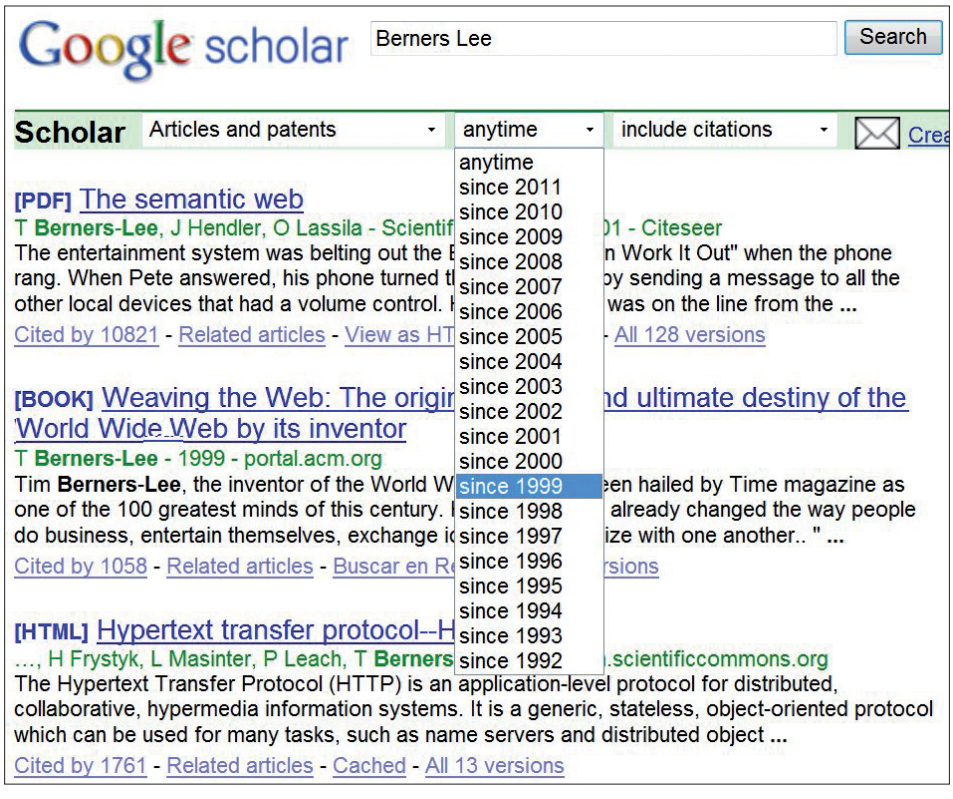
Sin embargo las ventajas se impusieron, ya que la competencia mejoró las prestaciones del *WoS* (ventanas de citación más amplias, nuevos indicadores) y su cobertura geográfica (con un cierto número de revistas no anglosajonas, sobre todo de ciencias sociales y humanas –que al parecer tienen un significativo menor impacto–). A medio plazo, trabajos de fusión de ambas bases de datos proporcionarán una mejor idea de las bondades y limitaciones de cada una de ellas, pero mientras tanto, cabe esperar la multiplicación de estudios disciplinares y/o temporales que remeden los ya realizados previamente con ayuda de *WoS*.

En ese contexto apareció un nuevo e interesante actor, *Google Scholar*, la base de datos académica del famoso buscador[4].

Dentro de la estrategia global de *Google* de recolectar toda la información posible[5], la dificultad de indizar la llamada internet invisible motivó la elaboración de un producto que no dependiera de los robots automáticos.

"La opacidad de *Google* respecto a las fuentes que utiliza ha dificultado el análisis global del buscador académico"

Scholar se nutre de una serie de acuerdos con productores y distribuidores de bases de datos



académicas y científicas de todo el mundo que ceden sus registros bajo distintas condiciones (tanto la lista de suministradores como los detalles de los contratos son secretos comerciales de *Google*).

Google proporciona ciertos valores añadidos (citas, enlaces, etiquetas) además de añadir la gigantesca sección académica de la web visible que aparece en el buscador general.

El resultado es una gran base de datos bibliográfica multidisciplinar que incluye citas a los diferentes artículos (fundamentalmente como ayuda a la recuperación). Es decir, es el tercer gran sistema de citas junto con *WoS* y *Scopus*, con la ventaja de su mayor tamaño y el hecho fundamental de ser de acceso gratuito. Se trata de un producto todavía en versión beta (¡desde 2004!), cuyo futuro no está garantizado y que al parecer es mantenido por un equipo muy reducido. Todo ello podría explicar la falta de normalización documental, muy necesaria en un producto multifuente tan heterogéneo formal y sustantivamente.

A pesar de los distintos problemas documentales de *Google Scholar*, la reciente aparición del software gratuito *Publish or perish*[6], que permite la captura directa de los registros y calcula automáticamente diversos indicadores (incluyendo distintas variantes del índice h), ha renovado y generalizado el interés por *Scholar* en la comunidad bibliométrica.

En la bibliografía de esta nota figura una selección de artículos que tratan fundamentalmente dos áreas: la comparación directa de *Google Scholar* con las otras grandes bases de datos de citas (*WoS* y *Scopus*), y la utilización de registros de *Scholar* para la realización de estudios bibliométricos.

Los árboles no dejan ver el bosque

La opacidad de *Google* respecto a las fuentes que utiliza (y la evolución temporal de dicha cobertura, que parece se incrementó significativamente en los últimos años) ha dificultado el análisis global del buscador académico. De hecho, el diseño de muchos estudios comparativos implicaba utilizar básicamente instituciones y autores de reconocido prestigio, para los que se obtenía una cierta equivalencia con los resultados obtenidos en los productos de "calidad contrastada" (basados más o menos en núcleos de Bradford). Las diferencias en los estudios disciplinares se atribuían a diferencias de cobertura y, en fin, otras discrepancias se atribuían a problemas y limitaciones técnicas que se trataban de describir y evaluar o simplemente se citaban sin más, como pretexto.

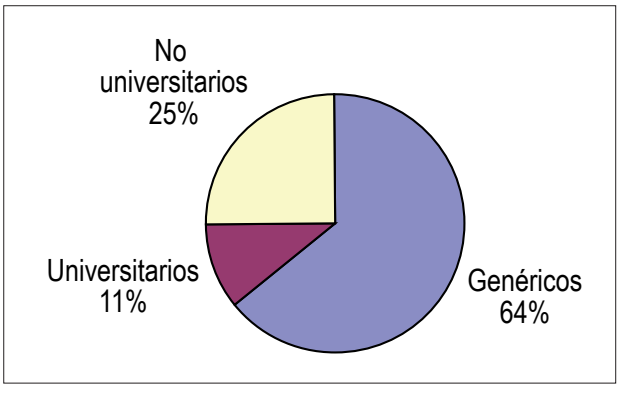
En el curso de un estudio cibermétrico sobre la distribución institucional de los contenidos recogidos en *Google Scholar*, descubrimos que las discrepancias son mayores de lo que se estimaba y que de hecho esta base no es comparable a *WoS* o *Scopus*, y su uso bibliométrico puede estar desaconsejado como norma general.

Se recogieron los registros totales (al menos con resumen) que aparecen en *Scholar* para dos grupos de dominios: 225 *top level domains* (incluyendo dominios nacionales como .es, .fr o .it, y los genéricos tales como .com, .org o .net) y 10.442 dominios universitarios (por ejemplo: ucm.es, harvard.edu u ox.ac.uk).

"*Google Scholar* es el tercer gran sistema de citas junto con *WoS* y *Scopus*, con la ventaja de su mayor tamaño y ser gratuito"

De la primera población se obtuvo un total de 86 millones de registros, de los que 55 millones (el 64%) correspondían a dominios genéricos, lo que cabría esperar de productores y distribuidores comerciales (.com) u organizaciones sin ánimo de lucro fuertemente presentes en este "mercado" (.org). Hay que tener en cuenta que *Google Scholar* muestra registros únicos, que "unifica" duplicados, es decir registros que pueden aparecer en repositorios institucionales o páginas personales pero que están también recogidos en distribuidores comerciales.

El segundo grupo (universidades) proporcionó 9 millones de registros, que supone un 10,6% del total obtenido en la estimación global de 86 millones, lo que implica que hay alrededor de un cuarto de los contenidos que bajo bandera nacional (dominio propio) son provistos desde instituciones no universitarias (productores locales, centros de investigación, portales, bibliotecas y repositorios digitales).



Results

Papers:	97	Cites/paper:	2.84	h-index:	7	AWCR:	23.23
Citations:	275	Cites/author:	260.06	g-index:	16	AW-index:	4.82
Years:	41	Papers/author:	52.14	hc-index:	5	AWCRpA:	21.50
Cites/year:	6.71	Authors/paper:	2.70	hI-index:	6.13	e-index:	13.82
				hI,norm:	6	hm-index:	6.50

Cites	Per year	Rank	Authors	Title	Year
101	8.42	1	L Codina	Evaluación de recursos digitales en l…	2000
65	5.42	2	L Codina	El libro digital y la WWW	2000
29	1.81	3	L Codina	El llibre digital	1996
18	1.00	4	L Codina	Modelo conceptual de un sistema d…	1994
11	1.83	5	L Codina	El libro digital y el territorio de la lec…	2006
8	0.47	7	LL CODINA	Teoría de recuperación de informaci…	1995
8	0.00	8	L Codina…	Web y cine: análisis comparativo de…	
5	0.00	13	L Codina	El nou sector emergent dels bancs …	
3	0.16	16	L CODINA	Software a prueba: Windows perso…	1993
3	0.14	15	LL Codina	Introducción a las bases de datos d…	1991

Resultado parcial del análisis de la producción del autor Lluís Codina mediante el programa Publish or Perish, http://www.harzing.com

La muestra universitaria puede utilizarse para un análisis en más profundidad, aunque hay que advertir que en muchos casos se trata de producción hospedada, es decir, además de trabajos publicados por personal de la institución se pueden encontrar contribuciones de terceros, tales como presentaciones en congresos celebrados en la universidad hospedadora o material didáctico producido por otros autores pero puesto a disposición (posiblemente sin cobertura contractual) por el profesorado propio.

En dicho análisis aparecen las sorpresas, ya que tras EUA, los siguientes países mejor representados son respectivamente España, Brasil y Taiwán (por delante de Japón, Alemania, Canadá y Reino Unido). Entre los veinte primeros aparecen también Costa Rica, México e Indonesia.

Descendiendo a nivel institucional, tras *Harvard* (base de datos de astronomía) se encuentran *Pennsylvania State University* (*CiteSeerX*), la *Universidad de La Rioja* (*Dialnet*), *Johns Hopkins University* (*MUSE*), *Catie* (Costa Rica, base de datos de agronomía), *Universidad Complutense de Madrid* (*CompluDoc*) o la *Universidad Autónoma del Estado de México* (*Redalyc*).

Es decir, de acuerdo con las actuales políticas institucionales, sus páginas web buscan reflejar no sólo la producción de "excelencia" de la universidad, sino todos los resultados independientemente de su calidad y tipo, e incluso hospedando producción de terceros, ya sea puntualmente o exhaustivamente como parte de consorcios amplios. *Google Scholar* está recogiendo y reflejando todo ello (y cada vez más, a medida que las iniciativas *open access* van triunfando, aunque sea lentamente).

En resumen, *Google Scholar* es una interesante herramienta de recuperación de información, con limitaciones derivadas de su falta de control documental, que se pueden soslayar dado su tamaño y el hecho de ser gratuita. La oferta de citas bibliográficas claramente incrementa su valor, pero la evolución reciente la aleja cada día más de aquellas que filtran contenidos de acuerdo con criterios de calidad (¿o impacto?). Este ruido extra desaconseja su uso liberal en los estudios bibliométricos, especialmente aquellos que tengan fines evaluativos.

"*Scholar* es una interesante herramienta de recuperación de información, pero se desaconseja su uso liberal en los estudios bibliométricos"

Notas

1. *ISI* (*Institute for Scientific Information*) es la empresa que en 1960 fundó **Eugene Garfield**, creador de las 3 bases de datos *Science Citation Index*. Fue comprada por *Thomson Reuters* en 1992.

2. *Dialog*, fundada por **Roger K. Summit** en 1980, fue comprada por *Thomson* en 2000, y revendida a *ProQuest* en 2008.

3. *WoK* (*Web of knowledge*) es el nombre comercial de un paquete de bases de datos de *Thomson Reuters* que incluye *WoS* (los 3 *citation indexes*), *Journal Citation Reports* (*JCR*), *Biosis*, *Derwent*, y otras.

4. *http://scholar.google.com*

5. *http://www.google.com/corporate*

6. *http://www.harzing.com/pop.htm*

Referencias bibliográficas

Bar-Ilan, Judit. "A closer look at the sources of informetric research". *Cybermetrics*, 2009, v. 13, paper 4. *http://www.cindoc.csic.es/cybermetrics/articles/v13i1p4.pdf*

Bar-Ilan, Judit. "Citations to the 'Introduction to informetrics' indexed by *WoS*, *Scopus* and *Google Scholar*". *Scientometrics*, 2010, v. 82, n. 3, pp. 495-506. DOI: 10.1007/s11192-010-0185-9.

Bar-Ilan, Judit. "Which h-index? A comparison of *WoS*,

Scopus and *Google Scholar*". *Scientometrics*, 2008, v. 74, n. 2, pp. 257–271. DOI: 10.1007/s11192-008-0216-y.
http://sci2s.ugr.es/hindex/pdf/Bar-Ilan2008.pdf

Beel, Joeran; **Gipp, Bela**. "Academic search engine spam and *Google Scholar*'s resilience against it". *Journal of electronic publishing*, 2010, v. 13, n. 3. DOI: 10.3998/3336451.0013.305.
http://quod.lib.umich.edu/cgi/t/text/text-idx?c=jep;view=text;rgn=main;idno=3336451.0013.305

García-Pérez, Miguel A. "Accuracy and completeness of publication and citation records in the *Web of Science*, *PsycInfo*, and *Google scholar*: a case study for the computation of h indices in psychology". *Journal of the American Society for Information Science and Technology*, 2010, v. 61, n. 10, pp. 2070-2085. DOI: 10.1002/asi.21372.

Harzing, Anne-Wil; **Van-der-Wal, Ron**. "A *Google Scholar* h-index for journals: an alternative metric to measure journal impact in economics and business". *Journal of the American Society for Information Science and Technology*, 2008, v. 60, n. 1, pp. 41-46.

Harzing, Anne-Wil; **Van-der-Wal, Ron**. "Google Scholar as a new source for citation analysis". *Ethics in science and environmental politics*, 2008, v. 8, n. 1, pp. 61-73. DOI: 10.3354/esep00076.
http://www.int-res.com/articles/esep2008/8/e008p061.pdf

Jacsó, Peter. "*Google Scholar* revisited". *Online information review*, 2008, v. 32, n. 1, pp. 102-114.
http://www.cs.unibo.it/~cianca/wwwpages/dd/08Jacso.pdf

Jacsó, Peter. "Savvy searching. Pragmatic issues in calculating and comparing the quantity and quality of research through rating and ranking of researchers based on peer reviews and bibliometric indicators from *Web of Science*, *Scopus* and *Google Scholar*". *Online information review*, 2010, v. 34, n. 6, pp. 972-982.

Kousha, Kayvan; **Thelwall, Mike**. "Sources of *Google Scholar* citations outside the *Science Citation Index*: a comparison between four science disciplines". *Scientometrics*, 2008, v. 74, n. 2, pp. 273-294. DOI: 10.1007/s11192-008-0217-x.

Li, Jie; **Burnham, Judy F.**; **Lemley, Trey**; **Britton, Robert M.** "Citation analysis: comparison of *Web of Science*, *Scopus*, *SciFinder*, and *Google Scholar*". *Journal of electronic resources in medical libraries*, 2010, v. 7, n. 3, pp. 196-217. DOI: 10.1080/15424065.2010.505518.

Mayr, Phillip; **Walter, Anne-Kathrin**. "An exploratory study of *Google Scholar*". *Online information review*, 2007, v. 31, n. 6, pp. 814-830.
http://www.ib.hu-berlin.de/~mayr/arbeiten/OIR-Mayr-Walter-2007.pdf

Meho, Lokman I.; **Yang, Kiduk**. "Impact of data sources on citation counts and rankings of LIS faculty: *Web of Science* vs. *Scopus* and *Google Scholar*". *Journal of the American Society for Information Science and Technology*, 2007, v. 58, n. 13, pp. 2105-25. DOI: 10.1002/asi.v58:13.

Mikki, Susanne. "Comparing *Google Scholar* and *ISI Web of Science* for earth sciences". *Scientometrics*, 2010, v. 82, n. 2, pp. 321-331. DOI: 10.1007/s11192-009-0038-6.

Torres-Salinas, Daniel; **Ruiz-Pérez, Rafael**; **Delgado-López-Cózar, Emilio**. "*Google Scholar* como herramienta para la evaluación científica". *El profesional de la información*, 2008, v. 18, n.5, pp. 501-510. DOI: 10.3145/epi.2009.sep.03.

White, Bruce. "Examining the claims of *Google Scholar* as a serious information source". *New Zealand library & information management journal*, 2006, v. 50, n. 1, pp. 11-24.
http://muir.massey.ac.nz/bitstream/10179/571/5/GoogleScholar.pdf