Google Scholar como herramienta para la evaluación científica

Por **Daniel Torres-Salinas, Rafael Ruiz-Pérez** y **Emilio Delgado-López-Cózar**

Resumen: *Google Scholar es un buscador especializado en recuperar documentos científicos y en identificar las citas que éstos han recibido, convirtiéndose de esta forma en un competidor de otros índices de citas. Diversos estudios han tratado de valorar su capacidad como herramienta bibliométrica. Debido a este interés se hace una introducción a su uso y a sus ventajas e inconvenientes frente a Web of Science y Scopus. Primero se analiza su modo de recopilar información y las propiedades de su interfaz. A continuación se describen los resultados a los que da lugar el buscador. En tercer lugar se analiza la cobertura de fuentes de información y los diferentes tipos documentales que recoge. Se expone cómo esta cobertura provoca un universo de citación diferente al de otros productos. Finalmente se especifican sus problemas de normalización y se expone una serie de precauciones a la hora de usarlo como herramienta de evaluación.*

Daniel Torres-Salinas *es doctor en documentación científica y trabaja como técnico de gestión de la investigación en la Universidad de Navarra, donde realiza auditorías sobre la calidad y el impacto de la investigación. Asimismo es miembro del grupo EC3 (Evaluación de la Ciencia y de la Comunicación Científica) de la Universidad de Granada donde participa en diferentes proyectos.*

Rafael Ruiz-Pérez *es profesor de catalogación en la Facultad de Comunicación y Documentación de la Universidad de Granada y miembro del grupo EC3. Sus líneas de investigación y publicación están centradas en la evaluación de revistas científicas y en la mejora de sus aspectos normativos. Es uno de los promotores de In-Recs: Índice de Impacto de las Revistas Españolas de Ciencias Sociales.*

Emilio Delgado-López-Cózar *es profesor de metodología de la investigación en la Facultad de Comunicación y Documentación de la Universidad de Granada y miembro del grupo EC3. Sus líneas de investigación se centran en la evaluación de revistas científicas y de la ciencia, el estudio de la investigación en ByD, y la evaluación del rendimiento científico. Es uno de los promotores del índice In-Recs.*

Palabras clave: *Google Scholar, Google Académico, Web of Science, Scopus, Bibliometría, Indicadores bibliométricos, Citas, Publicaciones científicas.*

Title: *Google Scholar* as a tool for research assessment

Abstract: *Google Scholar is a search engine that specializes in scientific information and in the identification of the citations that academic papers receive, making it a strong competitor for other citations indexes. For this reason, several studies have attempted to evaluate its capacity as a bibliometric tool. Due to this interest, we present an introduction to its use and the advantages and disadvantages versus Scopus and Web of Science. First, its way of collecting information and features of its interface are analyzed. The following section describes the results that Google Scholar generates. Thirdly, we analyze the coverage of information sources and the different document types to be found, showing how this coverage universe offers different citations versus other products. Finally, we specify the standardization problems of Google Scholar and offer a number of precautions that must be taken into account when using Google Scholar as an evaluation tool.*

Keywords: *Google Scholar, Google Académico, Web of Science, Scopus, Bibliometrics, Bibliometrics indicators, Citations, Scientific publications.*

Torres-Salinas, Daniel; **Ruiz-Pérez, Rafael**; **Delgado-López-Cózar, Emilio**. "Google Scholar como herramienta para la evaluación científica". *El profesional de la información*, 2009, septiembre-octubre, v. 18, n. 5, pp. 501-510.

DOI: 10.3145/epi.2009.sep.03

1. Introducción

En 1998 internet asistió al nacimiento de uno de sus grandes hitos, el buscador *Google*, creado por **Sergei Brin** y **Larry Page.**

Desde entonces la historia del buscador es conocida ya que desbancó al resto de competidores convirtiéndose en el principal portal de acceso a la información y la verdadera puerta de entrada a internet. Pese a su carácter generalista, se ha convertido en una herramienta in-

Artículo recibido el 11-03-09
Aceptación definitiva: 24-08-09

sustituible en el campo académico ya que gran parte de la comunidad científica lo emplea casi de forma diaria y sistemática. Según **Friend** (2006), cerca del 72% de los profesores lo utiliza para la búsqueda de artículos, lo que evidencia su enorme penetración. *Google Inc*. consciente de su presencia en este sector de usuarios y del enorme volumen de negocio que supone la información científica, lanzó a mediados de noviembre de 2004 *Google Scholar* (en adelante *GS*) o *Google Académico*, con el fin de proporcionar acceso universal y gratuito a las publicaciones científicas.

Es un producto que, a diferencia de las bases de datos bibliográficas tradicionales, no vacía contenidos de revistas sino que rastrea sistemáticamente la Web siguiendo la misma filosofía que *Google* pero haciendo converger en una sola plataforma dos servicios. En primer lugar es un buscador de publicaciones científicas y, en segundo lugar, es un índice de citas que ayuda a conocer el impacto que las publicaciones tienen. Precisamente esta última propiedad es la que más interesa y la que lo convierte en una competencia directa de otros índices de citación como *Web of Science* (*WoS*), de *Thomson Reuters*, o *Scopus*, de *Elsevier*. Por estas funciones *GS* se presenta a priori como una aplicación ideal para realizar al menos tres tareas:

– Buscar el texto completo de un trabajo.

– Buscar la producción bibliográfica de un autor, de una revista o sobre un tema.

– Buscar las citas que recibe un trabajo (libro, artículo de revista, tesis, informe…).

"*GS* es un buscador de obras científicas pero también es un índice de las citas que reciben, convirtiéndose en un competidor de *WoS* y *Scopus*"

En esta última función radica el enorme interés que tiene en general para los científicos que desean conocer la visibilidad de sus trabajos, y en particular para que evaluadores de la ciencia y bibliómetras puedan suplir las carencias de *WoS* y *Scopus*.

Otra de las particularidades fundamentales es su gratuidad, marcando una distancia enorme con el resto de proveedores, y más si tenemos en cuenta el elevado precio de las licencias de las bases de datos. Un ejemplo: la licencia nacional de *WoS* que proporciona la *Fundación Española para la Ciencia y la Tecnología* (*Fecyt*) para las universidades y organismos de investigación nacionales tuvo un coste para el trienio 2005-2008 de 25 millones de euros. En cierta medida *Google*, mediante *GS*, está fomentando un acceso universal a la información científica y además está viendo favorecida esta política por el incremento de la presencia de publicaciones científicas en acceso abierto, lo que ha hecho que se haya convertido en el aliado perfecto del movimiento *Open Access*.

Ante este panorama *GS* empieza a emerger como una alternativa a las bases de datos que tradicionalmente se han empleado para los estudios cuantitativos de la ciencia. La comunidad bibliométrica le está prestando gran atención tratando de desvelar sus principales funciones. En la mayoría de los análisis realizados, bien a favor (**Harzing**; **Van-der-Wal**, 2008) o en contra (**Jacsó**, 2005a; 2008b), se intenta calibrar su idoneidad como herramienta de valoración de la actividad científica y concretamente su impacto.

Por ello presentamos una síntesis de las principales propiedades de *GS*, centrándonos en cómo usarlo, en fijar su cobertura real y en analizar las ventajas y limitaciones que posee para su uso bibliométrico.

2. Funcionamiento e interfaz de búsqueda

GS se basa como *Google* en un robot, *Googlebot*, que de forma sistemática rastrea los contenidos de la Web, en este caso la Web académica, recopilando la información colgada de distintos dominios institucionales pertenecientes a universidades, repositorios, páginas de revistas, bases de datos e incluso catálogos de bibliotecas.

Una vez identificadas las referencias o los documentos, éstos son indizados registrando su descripción bibliográfica e incluyendo además las citas bibliográficas cuando se ha localizado el texto completo.

Los formatos que indiza son los habituales en el campo académico como doc o ppt, pero destaca especialmente el pdf seguido del html, aunque también podemos encontrar documentos en postScript. Esta indización a texto completo puede no ser del todo cierta ya que en determinadas ocasiones *Google* solo indiza 101 KB de los sitios web y lo mismo ocurre, aunque con mayor tamaño, con los documentos en pdf; podemos encontrar documentos en pdf de cierto tamaño que no están indizados completamente (**Price**, 2004; **Jacsó**, 2005b). Si efectivamente la información y el contenido relevante, como pueden ser las citas bibliográficas, se sitúan tras el límite de indización, éstas se pierden y no pueden ser recuperadas.

La interfaz de *GS* está basada en la proverbial sencillez de *Google* de manera que no resulta compleja al usuario: la pantalla principal sólo presenta una caja de búsqueda donde podemos introducir los términos que deseamos. Como en *Google*, disponemos de una serie

de operadores que pueden ayudar a mejorar la pertinencia de la búsqueda: el operador "+" permite incluir palabras vacías, "OR" expandir las búsquedas, "filetype:" especificar el formato del documento, "-" eliminar una palabra, o el uso de comillas localizar una frase exacta. Con la opción de búsqueda avanzada, podemos realizar búsquedas por tres campos: autor, título de la revista y año de publicación.

"*GS* indiza diferentes fuentes de información y variados tipos documentales"

Junto a estos tres campos reseñados, aunque sólo disponible en la versión inglesa, *GS* presenta un filtro para limitar los resultados por 7 grandes áreas científicas. Podemos localizar otras prestaciones avanzadas en el menú *Preferencias de Google Académico* donde, por ejemplo, desde la opción "idioma de búsqueda" los textos pueden ser limitados a una lengua concreta. En líneas generales, la interfaz apunta al minimalismo y huye de la sofisticación a la que nos tienen acostumbrados otros productos; las opciones de búsqueda, a pesar de tratarse de información científica, son bastante limitadas sobre todo si las comparamos con otras bases de datos bibliográficas. Así, *WoS* cuenta con 12 campos de búsqueda diferentes (tema, título, autor, grupo, publicación, año, dirección, congreso, lengua y tipo documental, ID proyecto y entidad financiadora) y *Scopus* incluye hasta 17 campos diferentes (ISSN, DOI, primer autor, etc.). Dichas bases de datos tienen opciones de filtrado de documentos muy completas que incluyen año de publicación, tipos documentales o revistas pudiéndose además obtener informes bibliométricos de los resultados como ocurre en *WoS* con los *Citation Reports*.

Convendría subrayar como una de las limitaciones principales de la interfaz de búsqueda de *GS* la ausencia de una opción específica de búsqueda para la localización directa de las citas que ha recibido un trabajo o un autor en un modo similar al que, por ejemplo, encontramos en *WoS* con *Cited Reference Search*.

"Gracias a su exhaustivo rastreo de la literatura científica *GS* alumbra un corpus documental que de otro modo sería difícilmente recuperable"

3. Presentación e interpretación de los resultados de búsqueda

Una vez lanzada la búsqueda *GS* devuelve los resultados que considera más pertinentes pero hay que tener en cuenta que sólo podrán ser consultados los 1.000 primeros. El algoritmo que ordena estos resultados se sostiene sobre la misma filosofía que el conocido *PageRank*, basado en una premisa tomada del mundo académico, donde los trabajos más citados son también los más importantes, haciéndola extensible al mundo Web mediante los enlaces.

Sin embargo, se incluye una serie de modificaciones sobre *PageRank* para adaptarlo a propiedades y convenciones propias del mundo científico y académico. Así, a la hora de ordenar los resultados en *GS* pesan otros factores. Por ejemplo, se considera el número total de citas recibidas, la disponibilidad del texto completo, el autor y la publicación (**Maryr; Walter**, 2007). Una vez ordenados los resultados, se muestran en un modo similar al de *Google*, aunque tienen una lectura diferente. Resumiendo, podemos encontrar al menos tres tipos de resultados diferentes (tabla 1):

– Enlaces al trabajo a texto completo. En este tipo de resultado obtenemos un enlace directo a la publicación original a texto completo al pinchar sobre el título. Se identifica por una flecha verde y el formato del documento entre corchetes.

1. Resultado que nos dirige al documento original a texto completo

[PDF] ▶ **E-strategias** en la introducción y **uso** de las **TIC** en la universidad
JM Duart, F Lupiáñez - Revista de Universidad y Sociedad del conocimiento, 2005 - ddd.uab.cat
... 1 1 5 © Josep M. Duart y Francisco Lupiáñez, 2005 © FUOC, 2005 **E-strategias** en la introducción y **uso** de las **TIC** en la universidad Josep M. Duart ...
Citado por 27 - Artículos relacionados - Versión en HTML - Las 15 versiones

2. Resultado que nos devuelve una cita

[CITAS] **Aprender sin distancias**.
JM **Duart** - Nueva Revista de Política, Cultura y Arte, 2000
Citado por 15 - Artículos relacionados - Las 2 versiones

3. Resultado que nos dirige a la fuente del documento

La motivación como interacción entre el hombre y el ordenador en los procesos de ...
JM Duart - dialnet.unirioja.es
... | Ayuda. **La motivación** como **interacción** entre el **hombre** y el ordenador en los procesos de formación no presencial. Autores: Josep ...
Citado por 3 - Artículos relacionados - En caché

Tabla 1. Tipos de resultados que se pueden obtener de GS

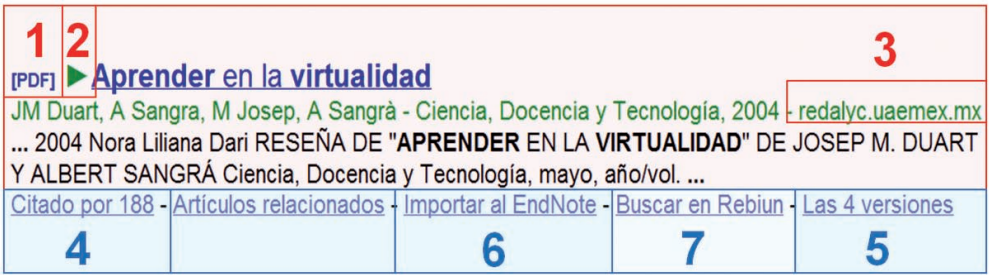

Figura 1. Elementos destacados de un resultado en GS

– Citas extraídas de documentos indizados. Los resultados vienen marcados con la etiqueta "[CITA]" y no presentan ningún tipo de enlace.

– Enlaces al documento en su fuente original. El resultado remite a alguna de las plataformas (repositorios y otras bases de datos) que *GS* rastrea. El acceso al documento depende de la plataforma.

"*GS* incluye indiscriminadamente todas las citas que es capaz de identificar, sin asegurar su calidad"

Todos los resultados, independientemente de su tipo, presentan una estructura similar (figura 1). En la zona superior encontramos una breve descripción bibliográfica del documento (título, autores, revista/fuente, año). Entre corchetes se indica ante qué documento nos encontramos, bien señalando el formato (pdf, html) o el tipo documental (libro, cita) (figura 1, n. 1). Se indica claramente si *GS* proporciona un acceso directo al documento con una flecha, situada en la zona derecha si el enlace conduce al texto original, o en la izquierda si redirige a otra fuente que proporciona el documento (figura 1, n. 2). También muestra cuál es el sitio web del que *GS* ha extraído la información (servidor, repositorio, etc.) (figura 1, n. 3); evidentemente esto no está disponible para el caso de "[CITA]".

En la parte inferior proporciona una serie de enlaces. Destaca en primer lugar "citado por", donde se muestra el listado de documentos recopilados por *GS* que citan el trabajo (figura 1, n. 4). Un segundo elemento interesante son las "versiones" (figura 1, n. 5) ya que agrupa bajo un mismo encabezamiento todas las versiones que ha localizado de un mismo trabajo, aunque no siempre realiza esta operación con precisión. Ejemplos de diferentes versiones de un mismo texto son los preprints, documentos de conferencias u otras adaptaciones, dándole a la versión del editor, si se indexa, el carácter de versión principal. La recopilación de las versiones facilita la agrupación de las citas dadas a un trabajo con independencia de su versión. Otras opciones de interés son la capacidad de exportar el registro a un software de gestión bibliográfica (figura 1, n. 6) o la posibilidad, si tenemos configurada esta función, de localizar el documento en una biblioteca gracias a la tecnología *Link Resolver* (figura 1, n. 7) (**Hartman**; **Mullen**, 2008).

"*GS* parece indexar cualquier revista, independientemente de su calidad"

4. Cobertura documental de *GS* y su impacto sobre la citación

Una de las propiedades que convierten a *GS* en un producto único e interesante es su amplia cobertura, que se pone de manifiesto con la indización de diferentes y variados tipos documentales (libros, informes científico-técnicos, *working papers* –informes de trabajo-, comunicaciones y ponencias en congresos, seminarios y jornadas, tesis y tesinas, etc.). Por tanto, no se limita a los trabajos publicados en revistas científicas, como en la mayor parte de las bases de datos. Normalmente el rastreador de *GS* toma sus registros de sitios donde la información se encuentra en libre acceso o de sitios comerciales que son procesados con el beneplácito de los editores, con los que previamente se ha llegado a algún tipo de acuerdo. Para entender la naturaleza de este producto mostramos algunos de los portales de información científica que cubre *GS* (**Jacsó**, 2005a; **Meho**; **Yang**, 2007):

– Repositorios: *arXiv.org*, *RePEc*, *E-Lis* o *CiteBase*.

– Portales de revistas: *HighWire Press*, *MetaPress*, *IngentaConnect*, *ACM Digital Library*.

– Bases de datos: *PubMed*.

– Editores comerciales: *Sage*, *Springer*, *Taylor & Francis*, *Nature, Blackwell*, *Macmillan*, *Wiley*, *Cambrigde University Press*.

– Sociedades Científicas: *American Physical Society*, *American Chemical Society*, *Royal Society of Chemistry*.

– Catálogos online de bibliotecas: *Worldcat*, *Dialnet*, *Institut de l'Information Scientifique et Technique* (*Inist*).

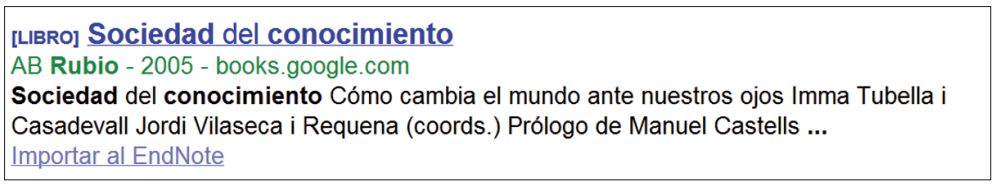

Figura 2. Búsqueda de un libro en GS remitiéndonos el enlace a Google Books

– Institutos y centros de investigación: *National Institutes of Health*, *NASA*, *American Institute of Physics*.

Por supuesto a estas fuentes habría que sumar los propios productos de *Google* como *Google Patents*, y sobre todo *Google Book Project* (figura 2), que ha escaneado ya más de un millón de ejemplares procedentes de 20.000 editoriales y bibliotecas en más de cien idiomas. Hay que señalar que el 10% de los mismos están escritos en español. Esto tiene mucha trascendencia ya que gran parte de los libros escaneados provienen de los fondos de bibliotecas académicas de las universidades del más alto prestigio como Standford, Princeton, Oxford, Harvard o Cornell o incluso de las colecciones de los servicios de publicaciones de las universidades como ocurre en el caso de la *Universidad de Salamanca*. Esta cobertura tan diversa de fuentes de información hace que podamos encontrar en *GS* una gran gama de tipos documentales:

– Libros

– Artículos en revistas

– Comunicaciones y ponencias a congresos

– Informes científico-técnicos

– Tesis y tesinas o memorias de grado

– Trabajos científicos depositados en repositorios o archivos de preprints

– Sitios web gubernamentales e institucionales

– Cualquier publicación con resumen

Quedan excluidos documentos no científicos como las reseñas de libros y editoriales, libros de texto, periódicos y revistas comerciales.

"Los datos de *GS* no tienen ninguna normalización, consecuencia de la amplia cobertura, la variedad de fuentes de información y el procesamiento automático de la información"

Una de las ventajas del exhaustivo rastreo de la literatura científica de *GS* es que alumbra un corpus documental antes casi invisible que de otro modo sería difícilmente recuperable (**Robinson**; **Wusteman**, 2007) al menos conjuntamente, y además permite hallar trabajos, sobre todo preprints, mucho antes de que aparezcan publicados en las revistas científicas comerciales (**Schroeder**, 2007). Sin embargo, el rastreo automático e indiscriminado conlleva también una importante limitación: muchos de los documentos indizados distan mucho del concepto de académico (**Noruzi**, 2005). No está claro qué entiende *GS* por "scholar" por lo que en ocasiones se incluyen entre sus resultados guías de bibliotecas, bibliografías de asignaturas o documentos administrativos. Esto se produce debido a que se suele indizar toda aquella información que cuelga de un dominio académico y el motor es incapaz de distinguir los tipos documentales propiamente científicos o académicos (**Friend**, 2006). Esta cuestión es importante ya que no tiene el mismo significado ser citado por un documento científico (libro, artículo, tesis…) que por otro que no lo es (programa de una asignatura…).

Asimismo, y esto es muy trascendente desde el punto de vista científico, aparecen mezcladas las citas provenientes de revistas arbitradas, es decir, las sometidas a *peer review*, con otras que no emplean ningún sistema de selección y evaluación de los manuscritos que publican. Para **Jacsó** (2008b) esta situación debería tenerse en cuenta a la hora de construir los indicadores bibliométricos ya que éstos tratan de medir el impacto científico a partir de fuentes de acreditada solvencia. Es evidente que *GS*, al incluir indiscriminadamente todas las citas que es capaz de identificar en cualquier documento, no puede asegurar ningún control de calidad de la información científica que presenta. Esta es la diferencia entre un entorno controlado (bases de datos tradicionales) y uno incontrolado (*GS*).

Independientemente de los errores que pueda cometer *GS* en el proceso de indización está claro que su cobertura documental genera un universo de citación diferente al de las otras bases de datos, con una serie de citas que son exclusivamente suyas. Algunos estudios han tratado de valorar el total de citas que puede aportar; por ejemplo **Kousha** y **Thelwall** (2007) sobre una muestra de 882 trabajos de diferentes áreas muestran como *GS* rescata 5.589 citas a los mismos, mientras que *WoS* recuperaba 4.184, con un solapamiento entre ambos de 2.387 referencias bibliográficas (es decir, el 24% del total de citas es común a ambas bases de datos –el 57% de *WoS* y el 43% de *GS*). Significa por tanto que *GS* recupera 3.202 citas únicas, aunque también pierde 1.797 respecto a *WoS*. Sin embargo, este solapamiento con *WoS* varía entre las diferentes áreas

científicas: en biología, física e informática gira en torno al 60%, mientras que en química se reduce al 33%. También **Meho** y **Yang** (2006) sobre 1.093 artículos de documentación compararon *GS*, *WoS* y *Scopus* determinando que entre las tres el total de citas únicas era de 5.288 (figura 3). *GS* localizaba un total de 4.184 mientras que las otras dos bases de datos conjuntamente sólo recuperaban 2.733. El solapamiento fue también del 24% (1.629/6.917).

Los datos por tanto parecen apuntar a que *GS* recupera un determinado número de citas únicas dependiendo de las disciplinas. Junto a esta situación, hay indicios además de que diversos tipos documentales podrían verse favorecidos con mayor citación como es el caso de los libros. **Harzing** y **Van-der-Wal** (2008) aportan algunas evidencias hacia una mejor cobertura de las citas recibidas por libros. Así, tomando las diez monografías ganadoras del *Terry Book Award*, *WoS* identificaba un total de 368 citas recibidas por estos libros mientras que *GS* elevaba la cifra hasta 783, lo cual supone un incremento del 128%. En definitiva, sabemos que *GS* no sólo es capaz de recuperar más citas sino que ofrece más citas únicas. Por ello es muy relevante conocer con cierta precisión el origen de las mismas desde una perspectiva documental.

Kousha y **Thelwall** (2008) en un estudio que recopila los trabajos publicados en 39 revistas de acceso abierto intentan revelar precisamente a qué tipos documentales corresponden las citas rescatadas por *GS* y que no son identificadas por otras bases de datos. Según los datos de estos autores, de un total de 5.589 citas, el 35% provenía de revistas científicas, el 25% de congresos/seminarios, un 22% de trabajos depositados

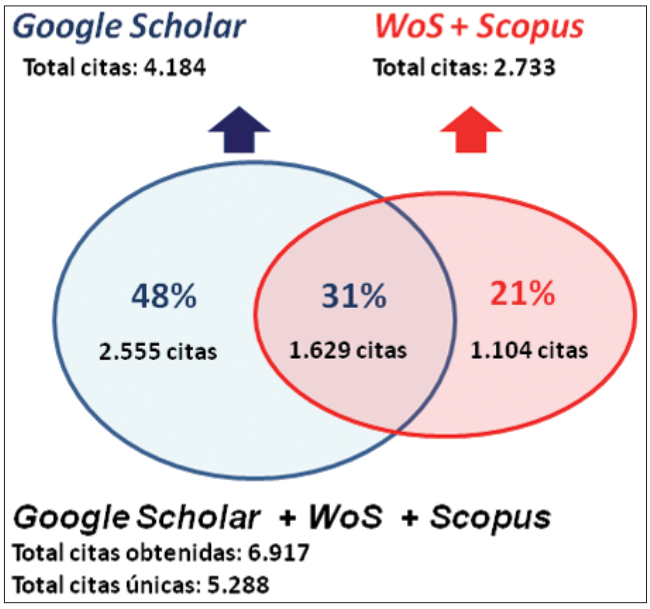

Figura 3. Ejemplo de Meho y Yang (2006) del solapamiento de las citas proporcionadas por GS, Scopus y WoS en el campo de la documentación

"Para la elaboración de un mismo ranking bibliométrico el procesamiento de los datos con *WoS* lleva 10 horas, con *Scopus* 20 y con *GS* 300"

en repositorios y, por último, un 8% de tesis doctorales. Matizan además que en función del área científica estos porcentajes varían. Así por ejemplo en física las citas recibidas de eprints/preprints llegan hasta el 48% y en informática las de congresos/seminarios se elevan al 43%. Estos datos son interesantes porque efectivamente reflejan que *GS* se adapta mejor a las prácticas de las distintas disciplinas, sobre todo aquellas que no utilizan como vía preferente de publicación las revistas científicas (humanidades, ciencias sociales, ingenierías).

En el trabajo de **Meho** y **Yang** (2006), referido a documentación, mientras que el 82% de las citas recuperadas por *WoS/Scopus* correspondían a artículos de revistas y el 18% restante a congresos, en *GS* eran del 43% y 34% respectivamente. Las tesis representaban el 10%, los libros el 6%, los informes el 5%, y otros documentos el 4%. Por tanto, las diferencias son claras: *GS* recupera citas de muy diversas fuentes, siendo las citas de revistas y libros, en porcentajes parecidos, las dominantes.

5. Cobertura de *GS* de revistas científicas

Pese a la capacidad de incorporar otros tipos documentales las revistas científicas siguen siendo el medio fundamental de comunicación por lo que conviene aclarar cuál es la cobertura de *GS* al respecto. Para lograr una perfecta cobertura de una revista, *GS* intenta llegar a acuerdos con editoriales como ocurre con *Science* o *Nature Publishing Group*, lo cual asegura una correcta indización.

No obstante, no siempre es así: *Elsevier*, la principal multinacional de la edición de revistas científicas en el mundo, se ha mostrado reticente ya que esta editorial es la que mantiene *Scopus*, que se puede considerar competencia directa de *GS* (**Meho**; **Yang**, 2007). Pero a pesar de ello, según **Bakkalbasi** et al. (2006) la mayor parte de los contenidos de *Elsevier* son indizados por terceros como el servicio *Ingenta*. Que la indexación de un grupo de revistas asociadas a un determinado editor dependa de un acuerdo es delicado ya que por ejemplo *GS* tampoco incluye las revistas de la *American Chemical Society* (*ACS*).

Esta situación provoca que determinadas disciplinas puedan presentar sesgos de bulto. En el caso de la química, que no esté un editor tan determinante como *ACS* provoca que de las citas recibidas por 373 artí-

culos publicados en estas revistas *GS* sólo sea capaz de rescatar 2.804 de un conjunto total de 8.723 citas recibidas (**Bornmann** et al., 2009). Un aspecto que conviene aclarar es que cuando un editor deja que *GS* incorpore sus datos no quiere decir que incorpore el texto completo de sus trabajos y la citación que generan los mismos. En la mayor parte de las ocasiones la información se reduce a proporcionar una mera referencia bibliográfica de los contenidos de las revistas.

En cualquier caso es interesante conocer cuál es la cobertura de *GS* respecto a otras bases de datos. Hemos de tener en cuenta que si éstas son empleadas como herramienta de evaluación es esencial conocer el universo de revistas empleado. En este sentido la política de *GS* es oscura ya que no proporciona ninguna información de cuáles son las revistas y qué tipo de indización tiene de cada una de ellas; tampoco sabemos a ciencia cierta cuáles son los editores que han firmado acuerdos con *GS*, información fundamental para conocer la validez de cualquier tipo de material científico (**Bauer**; **Bakkalbasi**, 2005).

Para solventar este problema las bases de datos bibliográficas presentan los denominados *Master List*, una información muy apreciada por bibliotecas y evaluadores. Ante la desinformación de *GS*, diversos trabajos han tenido como finalidad la comparación de su cobertura con la de otras bases de datos. Así por ejemplo **Mayr** y **Walter** (2007) estudiaron cuántas revistas de *WoS* están presentes en GS, determinando que del *Science Citation Index* (*SCI*) tiene el 85% (3.244) y del *Social Science Citation Index* (*SSCI*) el 88% (1.666).

Uno de los trabajos que mejor refleja la cobertura de las revistas científicas por parte de *GS* es el conducido por **Neuhaus** et al. (2006), que comprueba las revistas indizadas en 47 bases de datos de diversos campos. Estos autores indican que la cobertura de *GS* de diferentes disciplinas no es homogénea: cubre el 10% de las revistas de humanidades, el 39% de ciencias sociales, el 41% de educación, el 52% de economía y el 76% en ciencia y medicina. Por otra parte, en estos análisis se revela como *GS* tiene casi una cobertura total de las revistas en acceso abierto identificadas por diferentes directorios y bases de datos como el *Directory of Open Access Journals* (*DOAJ*) o *ACM Digital Library*. Asimismo *GS* incluye todas las revistas de *Pubmed* y *Pubmed Central* (**Neuhaus** et al., 2006).

Sin embargo, como hemos comentado con anterioridad, el hecho de que una revista esté presente en *GS* no significa que se permita el acceso al original ni que estén indizadas las referencias y citas de dicha revista; de ahí que *GS* pierda el 40% de la citación de revistas científicas de *WoS* y *Scopus* (**Meho**; **Yang**, 2007). Por último conviene señalar que mientras en *WoS* la selección de revistas se basa en un riguroso proceso de identificación de las más relevantes del mundo, *GS* no parece seguir ninguna directriz, por lo que tiene cabida cualquier revista, independientemente de su calidad.

6. El problema de la normalización y las búsquedas en GS

Consecuencia de la amplia cobertura, la variedad de fuentes de información empleadas y el procesamiento automático de la información es la ausencia de normalización en los datos de *GS*. Si en las bases de datos en general ya hay enormes limitaciones en la normalización de campos tan básicos como los autores o las instituciones, el problema cobra aquí mayores dimensiones. Si comparamos la información altamente estructurada de productos como *WoS* o *Scopus* podríamos decir que *GS* es un auténtico banco de datos tóxico que le resta credibilidad y le aleja de ser un producto consistente. Con el fin de orientar ante su posible uso evaluativo recopilamos algunos de sus errores más comunes y que aparecen bien documentados en la literatura científica sobre el tema (**Jacsó**, 2005a, 2005b, 2008a, 2008b).

El principal inconveniente de *GS* radica en que su herramienta de indización intenta detectar los campos que componen los documentos de forma automática, pero este proceso de identificación de estructuras no siempre funciona. En ocasiones se toman como autores de un trabajo elementos constitutivos del cuerpo del texto: por ejemplo, si desde la búsqueda avanzada introducimos como autor el texto "introducción" devuelve un total de 7.160 trabajos (figura 4). En este caso se toma el inicio de un epígrafe como autor. Una búsqueda por autor con "estado de la cuestión" da 1.330 resultados; igualmente ocurre con "índice" (4.320) o "contenido" (6.180). Inexplicablemente el campo autor presenta otros errores: el término "i-netlibrary" aparece en 12.200 ocasiones como firmante cuando el término ni siquiera aparece referenciado en los textos. El problema no sólo radica en una indexación automática sino en la absoluta ausencia de vocabularios controlados e índices (**Schroeder**, 2007). No hay siquiera un control de los títulos de las revistas (por ejemplo aparece tanto *BMJ* como *British Medical Journal*, *JAMA* y *Journal of the American Library Association*) ni de palabras ni términos clave como por ejemplo los *Medical Subject Headings* (*MeSH*) de *Medline* (**Shultz**, 2007), herramientas fundamentales para la recuperación pertinente de información científica.

También el campo del año de publicación provoca errores, y la búsqueda avanzada acotada por años genera resultados incomprensibles. En el momento de la realización de este trabajo si buscamos simplemente el período 2006-2008 *GS* devuelve un total de 93.900 documentos; sin embargo, al ampliar a 2005-2008 se

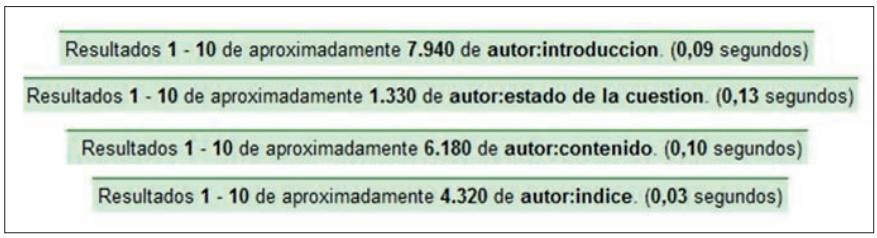

Figura 4. Problemas de indexación del nombre de los autores en GS

reduce a 89.800. Una búsqueda para el período 2004-2008 devuelve una cifra de 139.000 documentos, sin embargo una búsqueda entre 2000-2008 reduce los documentos a 109.000. Otro de los inconvenientes de las fechas es que a veces son identificadas erróneamente tomando como fecha de publicación del documento la fecha de depósito del mismo en un repositorio o incluso en otras ocasiones el número ISSN de las revistas. En general, cualquier número con cuatro dígitos es susceptible de ser confundido por *GS* con el año de publicación.

Otro de los inconvenientes encontrados es la enorme presencia de trabajos duplicados en los resultados, lo que crea confusión. Los duplicados se producen básicamente por las diferentes versiones que un artículo puede tener y que *GS* no ha sabido reagrupar bajo un mismo encabezamiento de título, por lo que aparecen como trabajos diferentes. Se pueden duplicar citaciones al estar presente la versión en preprint de un trabajo indizado en un repositorio y/o en la página personal del autor y la versión final publicada en una revista científica. Como consecuencia si comparamos la producción de una misma revista en *WoS* y en *GS* los resultados difieren bastante, situación que no se produce en otras bases de datos (*WoS* o *Scopus*). Si buscamos la producción en 2008 de una revista como *Lancet* *GS* devuelve 3.250 referencias, *Scopus* 1.653 y *WoS* 1.688. Es decir *GS* casi duplica los registros.

La consecuencia de estas incoherencias es un enorme aumento del coste en el tratamiento de datos derivados de *GS*. Para realizar un ranking bibliométrico (trabajos, citas, h-index) de 24 profesores del campo de la documentación, con *WoS* el procesamiento de los datos nos lleva 1 hora, 2 horas con Scopus y 30 horas si decidimos emplear *GS* (**Meho**; **Yang**, 2007). Esta situación pone en evidencia que, por el momento, su utilización a media y gran escala como herramienta de evaluación científica supone un consumo de recursos tan grande que la inhabilita. Sin embargo, ya empieza a haber algunas soluciones, como el software *Publish or Perish*, que facilita en alguna medida la recopilación y la manipulación de datos extraídos de *GS*. Este software ideado por **Harzing** proporciona indicadores asociados a los resultados, y hace posible la elaboración de rankings por diferentes campos y exportarlos a otros formatos como xls. La limitación principal radica en el elevado coste de limpieza de datos (normalización, eliminación de duplicados…) y en que no permite descargar las citas.

7. Consideraciones finales

Finalmente en la tabla 2, a modo de guía, se recopilan algunas de las particularidades de *GS* presentadas a lo largo de este trabajo frente a los índices de citas *WoS* y *Scopus*. En la misma se evidencia que *GS* es un producto ambicioso desde el punto de vista de su cobertura pero mal resuelto en el plano del procesamiento de la información y su presentación. En cualquier caso conviene señalar que ninguna base de datos tiene una cobertura completa de las citas que se emiten y cada una de ellas presenta un universo completamente diferente. Por esta situación los índices de citas disponibles en la actualidad son productos complementarios entre

Figura 5. Diferentes versiones de un mismo trabajo agrupadas por GS bajo encabezamiento de título único

Google Scholar		Índices de citas multidisciplinares (*Web of Science; Scopus*)	
PRECIO			
▲	Libre acceso	▼	Pago de licencias
COBERTURA GENERAL			
▼	Falta de transparencia en la cobertura. No se declaran acuerdos con editoriales ni las fuentes que se indizan	▲	Transparencia absoluta en las fuentes que componen las bases de datos. Disponibilidad de *Master Lists* actualizadas
▲	Cobertura de una amplia tipología de fuentes de información: repositorios, bases de datos, sociedades científicas, catálogos online de bibliotecas, institutos de investigación, productos de *Google* (*Google Patents* y *Google Books*)		
▲	Posibilidad de encontrar diversos tipos documentales: preprints, artículos de revistas, libros, tesis, informes, comunicaciones a congresos...	▼	Sólo cubren los contenidos de revistas científicas y recientemente libros de actas de congresos (*WoS*: *Conference Proceedings Citation Index*)
▼	Cobertura de documentos que podrían no ser de carácter académico: guías de biblioteca, temarios, etc.	▲	Contenidos exclusivamente científicos y mayoritariamente sometidos a revisión
▲	Buena cobertura de literatura en lenguas nacionales europeas	▼	Dominio de la literatura de carácter anglosajón, especialmente en *WoS*
▲	Acceso directo a publicaciones científicas a texto completo y gratuitas	▼	Acceso sólo a la referencia de los artículos
▲	Acceso directo al documento si la biblioteca lo tiene contratado	▲	Acceso directo al documento si la biblioteca lo tiene contratado
▲	Localiza citas emitidas por documentos no cubiertos por otras bases de datos, especialmente desde preprints, congresos o tesis doctorales. Esta característica lo hace especialmente útil para las siguientes disciplinas: humanidades, ciencias sociales e ingenierías	▼	Sólo localiza citas de revistas y congresos
COBERTURA DE REVISTAS CIENTÍFICAS			
▼	No existe ningún tipo de control en la selección de las revistas que indiza, por lo que todo tipo de revistas tiene cabida	▲	Rigurosos proceso de selección de las revistas científicas, especialmente en *WoS*
▼	Mala cobertura de las revistas de humanidades y ciencias sociales presentes en otras bases de datos (*MLA Bibliography*, *Philosopher's Index*, *PsycInfo*, *Sociological Abstracts*...)	▼	Tradicional mala cobertura de revistas de humanidades y ciencias sociales, aunque en la actualidad tienen una política de expansión en estos campos del conocimiento
INTERFAZ, BÚSQUEDAS y RESULTADOS			
▼	Sólo ofrece tres campos de búsqueda (autor, revista y año de publicación)	▲	Posibilidad de buscar en 12 campos diferentes en *WoS* y 17 en *Scopus*
▼	No tiene ninguna herramienta para analizar resultados	▲	Herramientas de análisis bibliométrico on-line como *Citation Report* en *WoS*
▼	Los resultados se presentan directamente ordenados y no existen otras opciones	▲	Permiten ordenar los resultados según diferentes opciones (título, nº de citas, fecha de publicación, primer autor...)
▼	Sólo permite exportar los resultados, uno a uno, a un software bibliográfico	▲	Exportación de los resultados en diferentes formatos (RIS, txt, tabulados, etc.)
▼	Gran coste en el procesamiento de los datos, lo que hace difícil su uso en estudios de gran escala	▲	Procesamiento de la información con menores costes en horas
▼	Presenta gran variedad de resultados duplicados		
▲	Posibilidad de exportar los resultados a software de análisis de datos: *Publish or Perish*	▲	Posibilidad de exportar los resultados a software de análisis de datos: *Histcite*, *Refviz*, *NWB*, *BibExcel*
▼	Sólo se muestran los 1.000 primeros documentos recuperados en cada consulta	▲	Se pueden consultar todos los resultados que genera una búsqueda
▲	Localiza las diferentes versiones de un documento y las agrupa bajo un mismo encabezamiento de título.		
▼	No identifica ante qué tipo documental nos encontramos. Tan sólo identifica los libros	▲	Cada registro está clasificado en un tipo documental (artículo, revisión, carta, nota, recensión, etc.)
▼	Sólo incluye el filtrado por 7 disciplinas	▲	Incluyen diversas opciones de filtrado (disciplina, año, tipo documental) que permiten refinar las búsquedas
CONTROL DE LA INFORMACIÓN			
▼	No existe normalización de los autores.	▲	No existe normalización pero tienen herramientas para identificación de autores (*WoS*=Author Finder)
▼	Ausencia de cualquier de tipo vocabulario controlado. No existe control de las revistas científicas; éstas pueden aparecer indizadas de diferente forma	▲	Control absoluto de las revistas científicas

Tabla 2. Comparación de las principales características de GS con las bases de datos multidisciplinares WoS *y* Scopus

sí. Mientras que hay un cierto consenso entre la comunidad científica en el uso de *WoS* como herramienta de evaluación, *GS* se muestra por el contrario como un producto inmaduro. Por esta razón se desaconseja su utilización como única fuente de información para la evaluación de la ciencia, especialmente en trabajos de media-gran escala (instituciones, países).

"*GS* es un producto ambicioso desde el punto de vista de su cobertura pero mal resuelto en el plano del procesamiento de la información y la presentación de resultados"

Ahora bien, creemos que *GS* es útil a nivel micro, como ayuda a los autores e investigadores concretos en la búsqueda rápida, fácil y directa de documentos a texto completo, y en la identificación de citas a sus trabajos. Sobre todo es útil para la literatura no anglosajona, que es la peor controlada por los sistemas de información dominantes en el mundo de la ciencia, para las disciplinas que no emplean preferentemente las revistas como medio de comunicación (ingenierías, humanidades, ciencias sociales…) y para localizar citas a libros, tesis, informes y a artículos publicados en revistas secundarias no incorporadas a la llamada "corriente principal de la ciencia".

8. Referencias

Bakkalbasi, Nisa; **Bauer, Kathleen**; **Glover, Janis**; **Wang, Lei**. "Three options for citation tracking: *Google Scholar*, *Scopus* and *Web of Science*". *Biomedical digital libraries*, 2006, v. 3, n. 7.
http://www.bio-diglib.com/content/3/1/7

Bauer, Kathleen; **Bakkalbasi, Nisa**. "An examination of citation counts in a new scholarly communication environment". *D-Lib magazine*, 2005, v. 11, n. 9.
http://www.dlib.org/dlib/september05/bauer/09bauer.html

Bornmann, Lutz; **Marx, Werner**; **Schier, Hermann**; **Rahm, Erhard**; **Thor, Andreas**; **Daniel, Hans-Dieter**. "Convergent validity of bibliometric *Google Scholar* data in the field of chemistry. Citation counts for papers that were accepted by *Angewandte Chemie International Edition* or rejected but published elsewhere, using *Google Scholar*, *Science Citation Index*, *Scopus*, and *Chemical Abstracts*". *Journal of informetrics*, 2009, v. 3, n. 1, pp. 27-35.
http://lips.informatik.uni-leipzig.de/files/2009-0.pdf

Friend, Frederick. "*Google Scholar*: potentially good for users of academic information". *Journal of electronic publishing*, 2006, v. 9, v. 1.
http://eprints.ucl.ac.uk/1771/1/JEP_OA_GS.pdf

Hartman, Karen; **Mullen, Laura-Bowering**. "*Google Scholar* and academic libraries: an update". *New library world*, 2008, v. 109, n. 5-6, pp. 211-222.
http://eprints.rclis.org/13820/1/GSfinalupdate.pdf

Harzing, Anne-Wil K.; **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.
http://www.int-res.com/articles/esep2008/8/e008p061.pdf

Jacsó, Péter. "As we may search - Comparison of major features of the *Web of Science*, *Scopus*, and *Google Scholar* citation-based and citation-enhanced databases". *Current science*, 2005a, v. 89, n. 9, pp. 1537-1547.
http://www.ias.ac.in/currsci/nov102005/1537.pdf

Jacsó, Péter. "*Google Scholar*: the pros and the cons". *Online information review*, 2005b, v. 29, n. 2, pp. 208-214.
http://www.jacso.info/PDFs/jacso-google-scholar-pros-and-cons.pdf

Jacsó, Péter. "*Google Scholar* revisited". *Online information review*, 2008a, v. 32, n. 1, pp. 102-114.
http://www.jacso.info/PDFs/jacso-GS-revisited-OIR-2008-32-1.pdf

Jacsó, Péter. "The pros and cons of computing the h-index using *Google Scholar*". *Online information review*, 2008b, v. 32, n. 3, pp. 437-452.
http://www.jacso.info/PDFs/jacso-pros-and-cons-of-computing-the-h-index.pdf

Kousha, Kayvan; **Thelwall, Mike**. "*Google Scholar* citations and *Google* web/url citations: a multi-discipline exploratory analysis". *Journal of the American Society for Information Science and Technology*, 2007, v. 58, n. 7, pp. 1055-1065.

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.

Mayr, Philipp; **Walter, Anne-Kathrin**. "An exploratory study of *Google Scholar*". *Online information review*, 2007, v. 31, n. 6, pp. 814-830.

Meho, Lokman I.; **Yang, Kiduk**. "Multi-faceted approach to citation-based quality assessment for knowledge management". *En: World library and information congress: 72nd IFLA General conference and council, 2006.*

Meho, Lokman I.; **Yang, Kiduk**. "Impact of data sources on citation counts and rankings of LIS faculty: *Web of Science* versus *Scopus* and *Google Scholar*". *Journal of the American Society for Information Science and Technology*, 2007, v. 58, n. 13, pp. 2105-2125.

Neuhaus, Chris; **Neuhaus, Ellen**; **Asher, Alan**; **Wrede, Clint**. "The depth and breadth of *Google Scholar*: an empirical study". *Libraries and the Academy*, 2006, v. 6, n. 2, pp. 127-141.

Noruzi, Alireza. "*Google Scholar*: the new generation of citation indexes". *Libri*, 2005, v. 55, n. 4, pp. 170-180.
http://www.librijournal.org/pdf/2005-4pp170-180.pdf

Price, Gary. *Google Scholar documentation and large pdf files*. 2004.
http://blog.searchenginewatch.com/041201-105511

Robinson, Mary L.; **Wusteman, Judith**. "Putting *Google Scholar* to the test: a preliminary study". *Program*, 2007, v. 41, n. 1, pp. 71-80.
http://www.ucd.ie/wusteman/articles/robinson-wusteman.pdf

Schroeder, Robert. "Pointing users toward citation searching: using *Google Scholar* and *Web of Science*". *Libraries and the academy*, 2007, v. 7, n. 2, pp. 243–248.

Shultz, Mary. "Comparing test searches in *PubMed* and *Google Scholar*". *Journal of the Medical Library Association*, 2007, v. 95, n. 4, pp. 442-445.
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2000776

Daniel Torres-Salinas
Grupo Evaluación de la Ciencia y la Comunicación Científica (EC3), Centro de Investigación Médica Aplicada, Universidad de Navarra, Avda. Pío XII, 31008 Pamplona (España).
torressalinas@gmail.com

Rafael Ruiz-Pérez*; *Emilio Delgado-López-Cózar
Grupo Evaluación de la Ciencia y la Comunicación Científica (EC3), Departamento de Biblioteconomía y Documentación, Universidad de Granada, Campus Cartuja, 18071 Granada (España)
rruiz@ugr.es
edelgado@ugr.es