



Facultade de Humanidades e Documentación

TRABALLO DE FIN DE MESTRADO

**GOOGLE ACADÉMICO NO SEU APOXEO. EXISTEN
DISCIPLINAS AÍNDA FÓRA DO SEU ALCANCE? ESTUDO DE
COBERTURA NA ÁREA DE ARQUITECTURA**

**GOOGLE ACADÉMICO EN SU APOGEO. ¿EXISTEN
DISCIPLINAS TODAVÍA FUERA DE SU ALCANCE? ESTUDIO
DE COBERTURA EN EL ÁREA DE ARQUITECTURA**

**GOOGLE SCHOLAR AT ITS PEAK. ARE THERE DISCIPLINES
STILL OUT OF REACH? COVERAGE STUDY OF
ARCHITECTURE**

Autor/a: Marcos Yáñez Arca

Director/a: Juan Ramón Rabuñal Dopico

Curso académico: 2013-2014

MESTRADO EN CIENCIAS DOCUMENTAIS NO CONTORNO DIXITAL

RESUMEN	4
1.- INTRODUCCIÓN.....	6
2.- ESTUDIOS DE COBERTURA DE GA.	9
2.1.- Introducción.....	9
2.2.- Metodología.....	9
2.2.1.- Comando “site:”	10
2.2.2.- Estudios de cobertura en base a búsquedas temáticas.....	11
2.2.3.- Estudios de cobertura a nivel título de revista	13
2.2.4.- Estudios de cobertura a nivel registro basados en una lista de referencia	14
2.2.5.- Estudios de cobertura basados en análisis de citas.	15
2.2.6.- Propuesta metodológica.....	17
2.3.- Resultados de los estudios. Estado de la cuestión.....	17
2.3.1.- ¿Existe un sesgo de GA hacia las ciencias “duras” en GA?	18
2.3.2.- Cobertura de GA en Ciencia, Tecnología y Medicina.....	20
2.3.3.- Ciencias Sociales.....	23
2.3.4.- Arte y Humanidades.....	24
3.- ESTUDIO DE COBERTURA EN EL ÁREA DE ARQUITECTURA	24
3.1.- Metodología.....	24
3.1.2.- Periodos de estudio dentro de la base de datos Avery	25
3.1.3.- Selección de la muestra.	25
3.1.4.- Comprobación de registros en GA.....	26
3.1.5.- Tipos de registros.....	26
3.1.6.- Fuentes de GA.....	27
3.1.7.- Estudio por países.	28
3.1.8.- Búsqueda en Google.	28
3.2.- Resultados.....	28
3.2.1.- Resultados de cobertura para los tres períodos.....	28
3.2.2.- Tipo de registros y fuentes.....	30
3.2.3.- Estudio por países (período 2010-2014).....	39
3.2.4.- Búsqueda en Google	41
3.3.- Discusión	41

3.3.1.- Editoriales STM y bases de datos.....	41
3.3.2.- Indexación de pequeñas editoriales por GA.	42
3.3.3.- Evaluación científica. El movimiento Open Access (OA) y las redes sociales científicas.....	44
3.3.4.- Análisis de los registros de GA sin enlace externo ni citas recibidas.	46
4. CONCLUSIONES.	49
BIBLIOGRAFÍA.....	52
ANEXOS	59
ANEXO I	59
ANEXO II	62
ANEXO III	64
ANEXO IV	69
ANEXO V	71

RESUMEN

Primeramente, se lleva a cabo un análisis sobre la metodología utilizada en más de medio centenar de estudios de cobertura del buscador científico Google Académico (GA). Para ello, se presenta una clasificación de los distintos enfoques y se detectan los principales problemas de aplicación sobre el buscador científico. En base a las conclusiones obtenidas se elabora una propuesta metodológica, cuyo resultado es una versión enriquecida del trabajo de Neuhaus et al. (2006). Posteriormente, se expone el estado de la cuestión sobre la extensión y profundidad del índice de GA para las distintas áreas científicas. En la segunda parte del trabajo se realiza un análisis de cobertura en el área de Arquitectura a partir de la metodología propuesta. Para ello se toma como referencia la base de datos Avery. Como resultado, se encontró que GA presenta tasas de cobertura muy bajas – aunque crecientes – para el área de Arquitectura en los tres períodos estudiados: 19,53% (1950-1989), 30,99% (1990-2009) y 36,03% (2010-2014). También se encontró una baja calidad de los registros, así como un desequilibrio de cobertura importante por países. Se discuten las causas y las posibles tendencias futuras.

Palabras clave: Estudios de cobertura; Google Académico; Arquitectura; Análisis de citas; Metodología.

ABSTRACT

Firstly, an analysis is carried out on the methodology used in more than fifty coverage studies of the scientific search engine Google Academic (GA). To this end, a classification of the different approaches is presented and the main problems of application to the scientific search engine are detected. Based on the conclusions obtained, a methodological proposal is elaborated, the result of which is an enriched version of the work of Neuhaus et al. (2006). Subsequently, the state of the question on the extension and depth of the GA index for the different scientific areas is presented. In the second part of the work, an analysis of coverage in the area of Architecture is carried out based on the proposed methodology. For this purpose, the Avery database is used as a reference. As a result, GA was found to have very low - but increasing - coverage rates for the Architecture area in the three periods studied: 19.53% (1950-1989), 30.99% (1990-2009) and 36.03% (2010-2014). Low quality records were also found, as well as a significant coverage imbalance by country. Causes and future trends are discussed.

Keywords: Coverage studies; Google Academic; Architecture; Citation analysis; Methodology

1.- INTRODUCCIÓN

El próximo noviembre Google Académico (GA) cumplirá 10 años y es posible que este aniversario le llegue en su mejor momento. Su éxito entre los miembros de la comunidad científica y académica es indiscutible como lo demuestran estudios recientes (Inger & Gardner, 2012). Scirus, uno de sus principales rivales, ha cerrado sus puertas el pasado Enero al dejar de ser “competitivo”, <https://www.facebook.com/ElsevierConnect/posts/491709367592128>, mientras su otro gran rival, Microsoft Academic Search, parece pasar por momentos de tribulación (Orduna-Malea, 2014). A su vez, GA no deja de presentar nuevos servicios: “Google Scholar Citation” (julio 2011), “Google Scholar Metrics” (abril 2012) o “Google Scholar Library” (noviembre 2013). Pesos pesados de la industria como Web of Science (WoS) firman con el buscador acuerdos de colaboración mutua (Kreisman, 2013). Y en revistas tan prestigiosas como Science se explica como sus funcionalidades han conquistado a científicos de la talla de Jonathan Eisen (Bohannon, 2014).

Los profesionales de la información y especialmente los bibliotecarios no han sido tan amables con GA en cambio, un recurso que ha alejado aún más a los estudiantes del catálogo, ha puesto en evidencia los nuevos proyectos de gestión de las colecciones (Chen, 2006) y pone en entredicho el valor de buena parte de las bases de datos suscritas por las instituciones (Chen, 2010a). Es por ello, que se han lanzado a un escrutinio sistemático del buscador científico. Estudios realizados han identificado una buena cantidad de errores en los metadatos de los registros de GA, en las citas recuperadas y en el funcionamiento de los operadores booleanos (Jacsó, 2005a; Jacsó, 2010). A esto se suma la falta de transparencia en la cobertura y el poco control sobre la calidad y autoridad de los trabajos que recupera (Gray et al., 2012). Más críticas ponen el foco en las bajas tasas de precisión que presenta (Giustini & Boulos, 2013); en las graves limitaciones en la sintaxis de búsqueda, en la presentación de los resultados y en la interfaz gráfica del usuario (Boeker et al., 2013); o en la facilidad para manipular citas y crear trabajos falsos (López-Cozar, Robinson-García & Torres-Salinas, 2014), etc.

Uno de los temas que ha provocado una mayor cantidad de literatura durante estos años ha sido la evaluación de la extensión y profundidad de su índice, elemento

esencial para la toma de decisiones para el uso o recomendación de un recurso de información. GA guarda sobre este aspecto un alto secretismo, tanto sobre el tamaño de su base de datos, de la lista de revistas que indexa o como de las fuentes que su araña rastrea. En Grey et al. (2012) se cuenta con detalle cómo intentaron en varias ocasiones ponerse en contacto con los administradores de GA para obtener más información sobre el contenido y extensión de su base de datos. La respuesta siempre fue la misma, que se trata de un servicio gratuito y que toda la información está en la página de ayuda del buscador. Si bien desde su puesta en marcha en noviembre de 2004 se han publicado distintas listas con las posibles editoriales con las que GA habría llegado a algún acuerdo para la indexación de sus fondos (Notess, 2005; Giustini & Barsky, 2005) y las bases de datos que permiten el acceso a su araña, lo que se sabe realmente sobre el origen, extensión y contenido de la base de datos de GA se debe a los estudios de cobertura.

Tenopir en 1982 consideró que había dos formas básicas de estudiar la cobertura de una base de datos. En la primera se usa una bibliografía o lista de publicaciones de referencia (“bibliography method”) para una disciplina con el objeto de testar una o varias bases de datos. De esta forma, la que presente un mayor grado de artículos recuperados tendrá una mejor cobertura.

El segundo método se basaría en el uso de una lista de términos con los que se pretende representar lo mejor posible una materia del conocimiento (“subject profile”). En este caso, los términos se lanzarían en distintas búsquedas en la base o bases de datos analizadas. La base de datos que más trabajos recuperase sería la que tendría una mejor cobertura (Rozeat, 2009).

Ahora bien, también se puede comparar la cobertura de GA en relación a otra base de datos a través de estudios de solapamiento, que permiten estudiar la semejanza de los contenidos y calcular el grado de información que pierde una fuente respecto a otra. Estos estudios se pueden llevar a cabo a su vez a nivel revista o a nivel registro, respecto a toda la población de la base de datos o respecto a una muestra (Pulgarín & Escalona, 2007). Para los estudios de solapamiento entre WoS, Scopus y GA normalmente se han utilizado análisis de citas.

El problema al aplicar estas técnicas sobre GA es que ésta no es una base de datos al uso, sino un buscador científico, lo que en cierta forma distorsiona las bondades

metodológicas comentadas. En cambio, no se ha encontrado ningún trabajo que tratase estas dificultades. Después de la revisión de más de 50 estudios de cobertura, en la primera parte del trabajo (punto 2) se propone una clasificación de las técnicas de estudio identificadas y se realiza un análisis crítico respecto a la validez de su aplicación a GA. Como conclusión a este estudio, se realiza una propuesta metodológica (apartado 2.2.6) basada en el modelo defendido por Chen (2010a) que a su vez toma como referencia el trabajo de Neuhaus et al. (2006). Posteriormente (apartado 2.3.), se presenta el estado de la cuestión sobre la cobertura de GA para las principales áreas del conocimiento, evaluando la validez de éstos una vez más en base a la metodología utilizada.

En la segunda parte del trabajo (punto 3) se realiza un estudio empírico para el área de Arquitectura en base al modelo propuesto en la primera parte del trabajo. Se toma como referencia la base de datos Avery, recurso fundamental en esta área del conocimiento.

A partir de tres muestras aleatorias con significación estadística (95% de confianza y 5 de intervalo) para los períodos 1950-1989 (pre web), 1990-2009 y 2010-2014, se determina la tasa de cobertura, se analizan los registros encontrados según criterios de accesibilidad y transparencia, y se clasifican e identifican las fuentes a partir de los que GA recupera los registros (portales de editores, bases de datos, páginas personales, etc.).

Además, para la muestra 3, se realiza un estudio para identificar posibles sesgos en el contenido de GA a favor de ciertos países o zonas geográficas (3.2.3) y se compara la capacidad de recuperación de los registros con su hermano mayor Google (3.2.4.). De esta forma, no sólo se trata de identificar el grado de cobertura de GA respecto al área de Arquitectura, sino también de conocer cómo construye su índice para esta disciplina, la calidad de los registros y la existencia de posibles desequilibrios y lagunas.

Finalmente, en el área de discusión (3.3) se analizan los bajos resultados de cobertura encontrados y se ponen en relación respecto a la tipología de las editoriales y a las características intrínsecas del área (3.3.1 y 3.3.2); se identifican nuevas tendencias que pueden permitir un mayor nivel de indexación (cambios en la evaluación científica, movimiento Open Access (OA) y redes sociales científicas) (3.3.3), para terminar

evaluando una tipología de registros, los registros de GA sin enlace externo ni citas recibidas (3.3.4), que se caracterizan por su baja calidad y nula accesibilidad al documento y que han presentado una incidencia notable para las tres muestras.

2.- ESTUDIOS DE COBERTURA DE GA.

2.1.- Introducción

Se han revisado más de medio centenar de trabajos que abarcan los 10 años de existencia del buscador científico. Para su identificación se ha realizado una serie de búsquedas en la base de datos LISTA (Ebsco) y GA, así como a través de la lectura detenidas de las bibliografías y apartados de revisión de los artículos. En este trabajo se reseñan no sólo los trabajos más citados, sino también aquellos que hacen referencia a disciplinas o áreas del conocimiento poco estudiadas.

En general, los artículos muestran un gran abanico metodológico lo que sin duda dificulta la homogeneización y comparabilidad de los resultados. También se ha identificado un fuerte desequilibrio en las áreas de conocimiento estudiadas, favoreciendo las Ciencias sobre las Ciencias Sociales y las Humanidades, así como una repetida comparación con las bases de datos Web of Knowledge (WoS) y Scopus en detrimento de bases de datos especializadas. Todo ello indica que aún es necesario realizar más investigaciones en esta línea de estudio.

La revisión bibliográfica se organizará aquí en dos grandes apartados. En el primero se estudian las metodologías aplicadas, donde se propone una clasificación de éstas y se analizan las principales implicaciones de cada una de ellas. Una vez estudiadas sus fortalezas y limitaciones, se podrá realizar una lectura general y comprensible de los resultados alcanzados por la investigación.

2.2.- Metodología

Como se verá en las siguientes líneas, el estudio de las técnicas utilizadas en los trabajos de cobertura es imprescindible para determinar la validez o la potencial generalización de los resultados obtenidos y de esta manera poder contextualizarlos y

graduarlos correctamente. A la teoría ya conocida sobre la evaluación de la cobertura de las bases de datos tradicionales, se debe añadir la variable GA, un buscador científico con características muy especiales que puede mermar la eficacia de las metodologías tradicionales. No se ha identificado ningún estudio previo en este sentido.

Para la realización del análisis se propone una clasificación de las distintas metodologías utilizadas desde una perspectiva puramente pragmática, adaptando en parte la ofrecida por Walters (2007):

- Trabajos basados en el uso del comando “site:” .
- Trabajos basados en búsquedas temáticas.
- Estudios de cobertura a nivel título de revista (uso del campo de búsqueda avanzada “Mostrar artículos publicados en”).
- Búsqueda título por título a nivel registro en relación a una lista de referencia.
- Análisis de citas.

2.2.1.- Comando “site:”

A lo largo del año 2005 Peter Jacsó publicó una serie de artículos de alto impacto en los que y a través de pequeñas búsquedas bien calculadas dejaba en evidencia las debilidades y disfunciones del recién creado buscador científico (2005a; 2005b). Jacsó considera que uno de los principales problemas de GA reside en la incapacidad de sus arañas de indexar satisfactoriamente las importantes fuentes a las que tiene acceso (los portales de las editoriales con los que ha llegado a acuerdos y las bases de acceso libre como PubMed o ArXiv). Para ello utiliza el comando <site:>. Este comando permite restringir la búsqueda a un servidor concreto. Así, el autor puede comparar los resultados obtenidos desde la base de datos nativa con los resultados obtenidos desde el buscador científico y determinar el grado de indexación de éste sobre aquel. De esta manera, Jacsó calcula que GA sólo era capaz de indexar una pequeña parte de Nature, de ADS (Astrophysics Data System) o de la revista Science (2005b). Giustini & Barsky utiliza la misma técnica para el portal del editor Blackwell y PubMed (2005).

Ahora bien, el comando <site:> no era en realidad una herramienta válida para conocer la capacidad de rastreo e indexación de GA. Pronto Chen (2010) desconfía de esta técnica, pero será en el 2012 cuando Arlitsch, & O’Brien desvelen finalmente el funcionamiento real del comando <site:> (2012). La clave reside en el proceso de “deduplicación” que el buscador lleva a cabo para fusionar en un solo registro las distintas

versiones de un trabajo (o copias rastreadas desde distintos servidores en la web). Una de ellas, normalmente la que da acceso al texto completo, se convertirá en la versión principal. Es aquella cuyo enlace será visible directamente al usuario. Con el comando <site:> sólo se recupera los registros de la copia seleccionada como principal, pero nunca de las otras versiones.



Ilustración 1: copia principal del registro procedente de Springer. Existen 7 copias en total del mismo trabajo en la base de datos de GA.

A pesar de los problemas metodológicos de estos trabajos, y que estas técnicas dejaron de aplicarse en beneficio de enfoques más sistemáticos, los resultados obtenidos en aquellas primeras publicaciones han tenido una gran repercusión en la comunidad científica a lo largo del tiempo, hasta el punto que aún en 2013 había autores que seguían sorprendiéndose por cómo GA, accediendo tan sólo a poco más de 1 millón de registros de PubMed, podía recuperar en cambio el 100% de la muestra estudiada (Gehanno et al.). La sombra de las técnicas de Jacsó es pues aún patente.

2.2.2.- Estudios de cobertura en base a búsquedas temáticas

Los análisis realizados sobre GA a través de búsquedas tuvieron su apogeo en los primeros años posteriores al nacimiento del buscador, años 2005-2006 principalmente, pero pronto fueron tachados como estudios poco serios. En este sentido Chen (2010b) decía lo siguiente:

No todos los análisis llevados a cabo hasta ahora han sido estudios empíricos basados en la búsqueda de grandes muestras de títulos. Algunas revisiones simplemente se basaban en búsquedas por una o pocas palabras clave y en la comparación del número de resultados devueltos por el sistema y las palabras clave contenidos en ellos. Desde que los diferentes sistemas de búsqueda pueden usar diferentes algoritmos de búsqueda y además no todas las búsquedas por palabra clave son relevantes, es natural que diferentes sistemas de búsqueda recuperen un distinto número de registros incluso buscando en la misma base de datos. Por tanto, es difícil evaluar la validez de este tipo de trabajos.

De todas formas, estos estudios pueden ser fiables si se llevan a cabo búsquedas lo suficientemente específicas. Por ejemplo, Notess (2005) realiza una búsqueda por <"protonation alkylation">: PubMed recupera 2.068 artículos, 1.524 Scirus y GA 1.820. La misma búsqueda en el portal del editor ACS (American Chemical Society) devuelve 21.685 registros. Jones (2005) lanza una búsqueda por <Nodilittorina>, un pequeño caracol marino en todas ellas. BIOSIS recupera 110 resultados y GA devuelve 102 trabajos con una tasa de cobertura del 27%, 39% y 56% para publicaciones posteriores a 1969, 1991 y 1996. En cambio, Gardner & Eng (2005) utiliza búsquedas excesivamente genéricas, <homeschooling OR "home schooling">, en GA, PsycInfo, ERIC y WoS, analizando los primeros 100 resultados para cada base de datos. PsycInfo devuelve 105 trabajos, pero GA devuelve 2.200. De esta forma, el trabajo deja fuera de análisis 2.100 trabajos. En ninguno de los tres trabajos anteriores se realizó un estudio cuantitativo de la relevancia de los resultados obtenidos, ni se buscó los trabajos título a título en GA para saber si realmente aparecían en la base de datos del buscador.

Un trabajo que deja palpables las limitaciones de esta metodología en los estudios de cobertura es el realizado por Levine-Clarck & Kraus en 2007. La investigación, en la que se compara a GA con la base de datos Chemical Abstracts Service (CAS), se lleva a cabo en dos fases. En la primera, la comparación se realiza a través de 5 búsquedas realizadas en ambas fuentes (2 temáticas, dos por componentes y una por autor). En la segunda fase se buscan todos los trabajos recuperados previamente título a título en ambas bases de datos. Las tasas de recuperación así obtenidas son ostensiblemente diferentes (ver tabla).

	Por tópico		Componente		Autor personal	
GA	84,1%	96,3%	26,1%	51,1%	52,7%	57%
CAS	19,8%	78,3%	77,8%	96,7%	97,8%	97,8%

Tabla 1: diferencias en las tasas de recuperación para la fase 1 y 2 en Levine-Clarck & Kraus (2007).

Como se puede observar la tasa de cobertura para la búsqueda temática se incrementa 15 puntos porcentuales en el caso de GA y más de 58 para PubMed cuando ésta se basa en una comprobación título a título. En la búsqueda por componente las diferencias entre la primera y segunda fase son de 25 y 19 puntos. Estos datos parecen invalidar totalmente una metodología basada en búsquedas para la realización de análisis de cobertura.

En cambio las búsquedas temáticas han mostrado ser de utilidad para otro tipo de investigaciones respecto a GA, como por ejemplo para analizar la calidad de los resultados devueltos por el buscador (Howland et al., 2009), determinar la tasa de recuperación y de precisión relativas (Walters, 2009 y 2011) o evaluar las posibilidades y limitaciones de la búsqueda avanzada (Boeker et al., 2013).

2.2.3.- Estudios de cobertura a nivel título de revista

Este tipo de estudios se basan en el uso del campo “Mostrar artículos publicados en...” de la búsqueda avanzada de GA el cual permite buscar trabajos a partir del título de una revista. Realmente este enfoque ha sido poco explorado hasta el momento, lo cual es plausible ya que se debe enfrentar a problemas metodológicos y prácticos importantes.

En relación a la primera dificultad, los estudios de cobertura a nivel revista entre distintas bases de datos dependen en buen grado de “la política de indexación de documentos que siga cada base de datos”, ya que mientras unas bases de datos pueden vaciar las publicaciones íntegramente otras pueden hacerlo de forma selectiva (Pulgarín Guerrero A., 2007). Esta objeción aún es más grave para el caso de GA donde las revistas son indexadas automáticamente, bien a través de los portales de los editores (en el mejor de los casos), de bases de datos con políticas de indexación más o menos arbitrarias, o incluso de forma indirecta a través de citas de terceros trabajos.

En relación a los problemas prácticos, éstos surgen por las graves limitaciones del campo de búsqueda de GA así como de la visualización de los resultados, a lo que hay que sumar la falta de control terminológico del buscador. En concreto, se podrían citar los siguientes:

- En un primer momento, el buscador no permitía el uso de expresiones exactas (comillas) dentro del campo de búsqueda por revistas (Jacsó, 2005b). Aunque actualmente GA sí permite utilizar comillas (“”), a partir de un número de resultados los problemas persisten al devolver artículos de publicaciones con nombres similares.
- El campo de búsqueda por título de revista permite un máximo de 60 caracteres. De esta forma puede haber problemas con la recuperación del contenido indexado respecto a revistas con títulos largos (Chen, 2010).
- La búsqueda por ISSN se muestra errática por lo que se descarta como forma de búsqueda (Harzing y Van der Wal, 2009)

- Al no existir un control de autoridades puede existir innumerables variantes del título de la publicación en el índice de GA.
- Imposibilidad de presentar los resultados de una búsqueda por título de revista ordenados por volumen y número, lo que dificulta seriamente los test de cobertura. (Norris, M., & Oppenheim, C., 2007).
- GA solo permite visualizar los primeros 1.000 registros. Esto invalida los análisis de cobertura por título de revista, al menos para aquellas publicaciones con un gran número de artículos indexados (Harzing A.-W., 2011).

Los trabajos de Mayr P. & Walter A.-K. (2007) y Salisbury L. & Tekawade A. (2006) dejan en evidencia este enfoque. En el primer caso se toma como referencia 9.500 revistas de Ciencias, Ciencias Sociales y Humanidades. Mientras que en el estudio a nivel revista GA presenta un nivel de indexación mayor para las revistas de Ciencias Sociales (88,11%) frente al área de Ciencias (85,2%), cuando se analizaron los registros de los artículos, se encontró que GA indexaba de una forma mucho más sistemática y consistente las revistas de Ciencias. En Salisbury L. & Tekawade (2006) para el área de Economía Agraria, se encuentra que a pesar que GA presenta la mayor tasa de indexación de revistas frente a EconLit y CAB Abstracts estas últimas recuperan más artículos. Por ejemplo, para el año 2005 GA indexa el 85,19% de las revistas mientras que CAB Abstracts indexa un 61,57. A pesar de ellos, GA sólo recupera 2.130 registros mientras que CAB abstracts recupera 3.550. Como indican los autores, mientras CAB Abstracts indexa las revistas número a número (“cover to cover”) la indexación de GA de éstas es muy irregular.

Como conclusión, los estudios a nivel revista en GA no son suficientemente consistentes y requieren ser acompañados con análisis a nivel registro.

2.2.4.- Estudios de cobertura a nivel registro basados en una lista de referencia

La fortaleza de esta metodología reside en su simplicidad ya que para llevar a cabo un estudio de cobertura de este tipo sólo se necesitaría seleccionar una lista de artículos y comprobar posteriormente su existencia en el índice de GA a través de la búsqueda de cada referencia título a título. Este enfoque es tan diáfano que un autor como Chen lo considera la metodología ideal para llevar a cabo estudios de cobertura. Eso sí, para ello se requiere contrastar el índice de GA con una base de datos especializada mediante el uso de muestras aleatorias (2010a). Los principales peligros de esta metodología se ciernen en no utilizar una lista de referencia suficientemente

representativa del área de estudio seleccionada y que el tamaño de la muestra sea tan pequeño que carezca de validez o significación estadística.

En el Anexo III se ha realizado un resumen de 11 trabajos de cobertura que han seguido esta metodología y se ha determinado en qué grado se cumplen los dos criterios expuestos por Chen.

2.2.5.- Estudios de cobertura basados en análisis de citas.

A pesar que los estudios infométricos basados en análisis de citas han sido utilizados para la evaluación de bases de datos, como por ejemplo a través de estudios de solapamiento (Bar-Ilan J., 2008), hay que tener en cuenta que no es lo mismo la cobertura de una base de datos como su capacidad para recuperar citas (Kousha K. & Thelwall M., 2008), lo que hace necesario para este tipo de estudios añadir una serie de procesos complementarios en el diseño metodológico.

A diferencia de los laboriosos trabajos de cobertura basados en el análisis a nivel registro, mediante la comprobación título a título de una lista de publicaciones, los análisis de citas permiten manejar una cantidad de datos mucho más extensa y con una considerable reducción en el consumo de tiempo gracias a aplicaciones informáticas como Publish or Perish (Harzing, 2011).

Además, y a diferencia de los estudios de cobertura basados en una lista de referencia o “golden standard database”, en este tipo de enfoques es habitual no sólo contabilizar el número de citas recuperado, sino también calcular la tasa de citas únicas para cada base de datos, así como la tasa de solapamiento. Por ello, ahora el análisis respecto a GA no es pasivo en relación a una base de datos de referencia, sino que también permite conocer aquellos contenidos que indexa GA, pero que no recogen las otras bases de datos del estudio. En este sentido, los análisis de citas suelen ofrecer una imagen más global sobre la cobertura de las distintas bases de datos.

Ahora bien, los análisis de citas realizados sobre GA se enfrentan con una gran cantidad de problemas técnicos basados en la poca calidad de su índice, en la falta de control de autoridades y en ciertas limitaciones del software. Así, el número de citas que GA muestra bajo el registro es sólo un indicador aproximado y no siempre coincide con el número real (Jacsó, 2006; Levine-Clark M. & Gil E., 2009); existen graves problemas para diferenciar autores con nombre similares (Harzing, 2013); algunos registros en GA aparecen duplicados debido a que se recogen distintas variantes del

título (por ejemplo con o sin subtítulo); existencia de citas falsas (Jacsó, 2006); no todas las citas recuperadas por GA tiene carácter académico, por ejemplo aquellas que proceden de bibliografías recomendadas (Jacsó, 2006; Harzing, 2013); GA no tiene acceso a las bibliografías de todos los trabajos que indexa por lo que muchas de las citas que no son recogidas en estos análisis realmente están incluidas en el índice (Meho & Yang calcularon una tasa de error del 12% (2007), mientras que en Bar-Ilan (2010) se incrementaba la tasa de cobertura de GA sobre Scopus y WoS de un 68,8% a un 94,74% cuando las citas se buscaron título por título); GA no permite la visualización de los registros más allá de los primeros 1.000 resultados lo que limita en gran medida la posibilidad de realizar análisis de citas para revistas o entidades agregadas (departamentos, universidades, etc.) (Harzing, 2011).

Debido a los problemas anteriores y tomando como modelo los trabajos realizados por Meho L.I. & Yang K. (2007) y de Bar-Ilan J. (2010), se puede considerar que para que un análisis de citas en GA permita una interpretación diáfana respecto a la cobertura de su base de datos, el estudio llevado a cabo debe seguir los siguientes pasos:

- Selección de una muestra de trabajos (obras citadas) representativa para el área de estudio.
- Búsqueda en GA de posibles duplicados o triplicados de las obras citadas. Identificación de todas las variantes del título existentes en el índice si las hubiere.
- Descarga de los registros de las obras citantes y de sus correspondientes enlaces. Recuento.
- Comprobación de los enlaces para determinar si realmente la publicación cita la obra objeto de estudio.
- Proceso de limpieza de registros (eliminación de citas de carácter no científico, registros inaccesibles y duplicados).
- Clasificación de cada obra citante dentro de su tipología documental.
- Cálculo de la tasa de solapamiento y citas únicas para las distintas bases de datos que forman parte del estudio. Cálculo por tipología documental.
- Comprobación de que las citas únicas no están indexadas en las otras bases de datos de estudio (búsqueda título a título).

2.2.6.- Propuesta metodológica.

Existen muchas áreas del conocimiento para las que todavía no se ha realizado estudios de cobertura respecto a GA y que por tanto se desconoce la extensión y profundidad de su índice. Algunos ejemplos son la Antropología, Ciencias Políticas, Urbanismo, Derecho, Filosofía, Arquitectura, etc. Para este tipo de estudios iniciales se propone un modelo enriquecido de los trabajos de Neuhaus et al. (2006) y Chen (2010a) que se puede sintetizar en los siguientes puntos:

- Selección de una o varias bases de datos referenciales respecto del área de estudio. Los resultados serán más relevantes cuanto más exhaustiva respecto a la disciplina sea la base de datos seleccionada.
- Se debe tratar de un trabajo longitudinal, es decir, que cubra el mayor período posible dentro de la base de datos de estudio. Para ello, puede ser interesante realizar varias muestras que abarquen distintas décadas u otras divisiones temporales, como se realizó en Meier & Conkling (2008).
- Las muestras deben tener un tamaño mínimo que permita una interpretación estadística de los resultados. Unos criterios mínimos serían un 95% de confianza con un intervalo del 5%.
- Además de conocer la tasa de cobertura, es fundamental conocer el grado de calidad y accesibilidad de los registros, lo que dará una idea de la profundidad del índice así como de la dificultad de recuperación de los registros en búsquedas reales. En este punto, un modelo a seguir es el trabajo de Christianson (2007).
- Identificación de fuentes. Esto permite estimar el grado de consistencia y exhaustividad con la que se indexan por ejemplo las revistas científicas y un mayor conocimiento de cómo GA construye el índice. En este punto se puede seguir a Mayr & Walter (2008).
- Finalmente, estudio del grado de cobertura según el idioma y área geográfica. De esta manera es posible identificar sesgos en el índice de GA así como posibles lagunas.

2.3.- Resultados de los estudios. Estado de la cuestión.

En este apartado se trata de ofrecer el estado de la cuestión sobre el grado de cobertura de GA para las distintas áreas de la ciencia. Esto no es una tarea sencilla. En primer lugar porque existen importantes lagunas de estudio en relación a ciertas áreas del conocimiento como por ejemplo las humanidades y las artes, o ciertas disciplinas de las

ciencias sociales como por ejemplo las ciencias jurídicas. Para todas ellas apenas se han podido encontrar indicios o pequeñas estimaciones. Pero también porque las técnicas de análisis encontradas son muy heterogéneas con distintos niveles de rigurosidad empírica o de tamaño muestral, lo que hace difícil encontrar resultados homogéneos y comparables entre las distintas disciplinas.

Para una mayor claridad expositiva se ha dividido este apartado en varias subsecciones, tratando de responder en cada una de ellas a una importantes cuestión sobre la cobertura de GA. Cuando se ha visto necesario, se ha indicado junto con los resultados las dudas existentes sobre la metodología utilizada.

2.3.1.- ¿Existe un sesgo de GA hacia las ciencias “duras” en GA?

Aunque GA nunca ha publicado la lista oficial de las editoriales con las que originalmente llegó a un acuerdo para la indexación de sus fondos, algunos autores han tratado el asunto con cierta prolijidad como por ejemplo M. Burright (2006). Incluso D. Giustini & E. Barsky publicaron en el “Apéndice A” de su artículo de 2005 una lista con los 29 participantes del proyecto piloto CrossRef, posible precedente de GA. Revisando los socios originales de GA a partir de estos trabajos se encuentran nombres de la talla de ACM, Blackwell, IEEE, Nature Publishing Group, Springer, Taylor & Francis, Wiley, Oxford University Press, Institute of Physics Publishing etc. A este grupo posteriormente se unirían editoriales como Elsevier (al menos parcialmente a través de la indexación del portal ScienceDirect), ACS y otras (Brantley, 2007; Harzing, 2013). Como se puede observar, en su gran mayoría se trata de editoriales STM (Science, Technology & Medicine).

Un patrón parecido se encuentra en las bases de datos de acceso libre. Consultando el “Ranking Web de Repositorios” elaborado por el CSIC (<http://repositories.webometrics.info/es>) y ordenando los resultados por tamaño, las fuentes que se encuentran en las primeras posiciones son bases de datos especializadas como PubMed (Medicina), el repositorio del CERN (Física), CiteSeerX (principalmente Informática y Ciencias de la Documentación), NASA Astrophysics Data System (Física, Astronomía...), Scielo (Salud), etc.; redes sociales científicas como Academia.edu y ResearchGate, principalmente especializadas en las áreas “duras” de la ciencia (entre un 30%-40% de los científicos registrados a ResearchGate son biólogos o médicos) y en menor medida en ciencias sociales; bases de datos de carácter multidisciplinar como

CNKI (China), HAL (Francia) o Dialnet (España); así como repositorios institucionales de composición temática indeterminada.

La idea de que GA ofrece una indexación más exhaustiva para el área de Ciencias que respecto a las áreas de Ciencias Sociales y Humanidades, al menos respecto a los artículos de revista, puede ser verificada o no dependiendo de la metodología utilizada. Así, Neuhaus comparó el contenido de GA respecto a 47 bases de datos para las principales áreas científicas y encontró que de media GA recuperaba el 76% de los contenidos en Ciencia y Medicina, mientras que sólo alcanzaba una tasa entre el 39 y el 41% para las Ciencias Sociales y del 10% para las Humanidades (2006).

La hipótesis, en cambio, no se verifica si se tiene en cuenta los estudios basados en análisis de citas. Bosman et al. (2006) encontró que mientras Scopus y Web of Science (WoS) recuperaban muchas más citas para las áreas de Química y Física, GA en cambio las superaba en las áreas de Ciencias Sociales y de Economía. Resultados parecidos, esta vez para una muestra mucho mayor, fueron obtenidos por Kousha & Thelwall un año después (2007). Mientras GA sólo recuperaba de media el 43% de las citas de WoS para Química y un 85% en Biología, sus cifras eran muy superiores para el área de Informática (201%), Psicología (200%), Sociología (219%), Educación (507%) y Economía (769%). Siguiendo con este tipo de trabajos, recientemente Harzing (2013) ha llevado a cabo un estudio longitudinal sobre 20 premios nobel de las áreas de Física, Química, Medicina y Economía. GA sigue recuperando menos citas en Química (aunque las diferencias se han reducido ostensiblemente), recupera ya más citas para las áreas de Física y Medicina y muchas más para el área de Economía (donde WoS pierde el 83% de las citas respecto a GA).

El análisis de los datos vistos en el párrafo anterior lleva a una falsa contradicción. Mientras Neuhaus comparaba el contenido de GA respecto a bases de datos especializadas, los estudios de citas comparan a GA con WoS y Scopus, bases de datos de citas multidisciplinares, pero con un sesgo muy importante a las áreas de Ciencias y Medicina. Por otra parte, no todas las disciplinas tienen las mismas prácticas respecto a las formas de publicación. Por ejemplo, Kousha & Thelwall (2008) mostraron la dependencia casi total de los químicos en las revistas científicas para la difusión científica, mientras que para el área de Informática y Física encontraron un gran peso respectivamente de los trabajos presentados a Congresos y de los preprints. Otro ejemplo es el caso de los libros. Analizando todos los trabajos presentados a la agencia

evaluadora de Inglaterra (RAE) para el año 2008, en el área de las Ciencias los libros de autor representaban escasamente el 0,4% de las publicaciones mientras que para el conjunto de las Ciencias Sociales alcanzaba el 12,38% (15,7% en Filosofía, 21,8% en Ciencias Políticas y el 23,9% en Historia) (Kousha et al., 2011). En cambio, WoS y Scopus son bases de datos que indexan fundamentalmente revistas científicas, es decir, no tienen en cuenta otras tipologías documentales (o en un grado anecdótico).

Como conclusión a este punto y retomando las ideas de Jacsó, un estudio bibliométrico de cobertura no debe tener en cuenta exclusivamente el tamaño del índice, sino también la composición de la base en términos de tipología documental, lenguas y cobertura geográfica, así como el rango de fechas cubierto por cada fuente (Rozear, 2009). Sin la determinación de estas variables, los resultados no son comparables y menos si éstas además se quieren extender a análisis entre distintas disciplinas con estrategias de difusión tan diversas.

2.3.2.- Cobertura de GA en Ciencia, Tecnología y Medicina.

Aunque como se ha visto GA contó desde el primer momento con un buen número de portales de grandes editoriales y bases de datos disponibles para ser rastreados por su araña, durante los primeros meses de existencia del buscador los resultados de cobertura, incluso contra bases de datos accesibles, fueron pobres, además de presentar problemas importantes de actualización (Jacsó, 2005b; Giustini & Barsky, 2005). De todas formas, a partir del año 2007 en adelante se encuentran mejoras sustanciales en la indexación, tanto gracias a una indexación más exhaustiva como al acceso a nuevas fuentes como los portales de ScienceDirect (Elsevier) y ACS (Chen, 2010b; Harzing, 2013).

2.3.2.1.- Ciencias de la salud

Se trata de una de las áreas en las que GA ha demostrado un mayor grado de cobertura desde sus comienzos. En Neuhaus et al., se encontró que GA era capaz de indexar el 100% de PubMed y el 94% de BioMed Central. En cambio su eficacia se reducía bruscamente al 46% en relación a CINAHL (principal base de Enfermería) (2006). Shultz (2007) hizo una comparativa entre GS y PubMed un año después a través de 10 búsquedas lanzadas en ambos recursos. A través de GA se encontraron 247 trabajos mientras que PubMed recuperó 147. Posteriormente se analizaron las 125 citas únicas de GA y se encontró que el 13% correspondían a tipologías documentales no cubiertas por PubMed y el 4,05% eran artículos pertenecientes a revistas no indexadas por la

base de datos. En el 2013, distintos trabajos basados en la búsqueda a nivel registro título por título encontraron que GA era capaz de recuperar el 100% (Gehanno et al., 2013), el 95% (Giustini & Boulos, 2013), el 92,9% (Boeker et al., 2013) y el 98% (Bramer et al., 2013) de las muestras analizadas.

De Groote & Raszewsky (2013) realizan un estudio interesante para el área de Enfermería comparando GA, Cinahl, Scopus y WoS a través del análisis de citas. Se tomó como muestra la producción científica de 30 investigadores de una Escuela de Enfermería de primer nivel del Medio Oeste (EEUU). Del total de los trabajos citados, Scopus recuperó 974 publicaciones, GA 927 y WoS 795. Respecto al análisis de citas llevado a cabo para 100 de estas publicaciones, Cinahl recuperó 966 citas (693 fueron citas únicas), WoS recuperó 1.406 citas (todas artículos de revistas), Scopus 2.437 y GA 3.497 (la mayor parte de ellas artículos de revista, libros y tesis de doctorado, máster y grado, aunque no se indicaron los porcentajes). A pesar del número de citas tan superior de GA, la tasa de cobertura respecto a las citas recuperadas por las otras 3 bases de datos fue del 45,4% (un porcentaje de todas formas que podría ser superior si se realizase una búsqueda título a título en el índice de GA)

2.3.2.2.- Ingeniería e Informática

Meir & Conkling (2008) realizaron un estudio de cobertura respecto a la base de datos referencial Compendex. La muestra de 960 registros se obtuvo de forma aleatoria a partir de los resultados de 8 búsquedas lanzadas repetidamente para cada una de las décadas desde 1950 hasta 2000. Como media, GA recuperó el 66% de los registros de Compendex, aunque los resultados ofrecieron una fuerte variabilidad desde el 33% y 51% para la década de los años 50 y 60, hasta una tasa de cobertura del 89% y del 88% para la década de 1990 y 2000 respectivamente. Este mismo estudio fue replicado 5 años después obteniendo una tasa de recuperación media del 75% de los registros, mejorando los resultados obtenidos en prácticamente todas las búsquedas.

En relación a los análisis de citas, varios estudios fueron realizados para el área de Ingeniería y de Informática. El problema para estas áreas es que los trabajos presentados a congresos tienen una gran tradición en la disciplina, tipología que WoS no indexan y que Scopus realiza sólo parcialmente. Como en la mayor parte de los estudios analizados no se realiza un estudio detallado por tipología documental, es difícil identificar con precisión las diferencias de contenido reales existentes entre las bases de datos. Bar-Ilan & Levene (2007), comparando las citas recuperadas en WoS,

Scopus y GA a partir de los 47 autores israelíes más citados dentro de la plataforma de Web of Science, encontraron que GA recuperaba muchos más trabajos citados y muchas más citas que las otras dos bases de datos. Por ejemplo, para el investigador en computación Oded Goldreich WoS indexaba 8 trabajos y recuperaba 326 citas, Scopus indexaba 5 trabajos y recuperaba 246 citas, mientras que GA recuperaba 39 trabajos y 2.957 citas. Para los otros autores del área de Informática e Ingeniería los resultados fueron similares. No se calcularon las tasas de solapamiento. Kousha & Thelwall (2008) en un análisis sobre 4 revistas de Informática, encontraron en WoS 1.117 citas (485 eran citas únicas) mientras que desde GA se recuperaron 2.884 (2.252 citas únicas). Como no se buscaron los registros título a título en GA, se desconoce la tasa de cobertura real de GA sobre las 1.117 citas encontradas en WoS.

2.3.2.3.- Otras disciplinas.

Pocas de estas áreas han sido estudiadas a través del modelo propuesto por Chen por lo que es difícil valorar la cobertura real de GA cuando los análisis de citas realizados no cumplen con procesos de calidad mínimos como la limpieza y verificación de citas, comparativas por tipología documental y rango de años, etc.

- Física. Los resultados entre WoS y GA son similares en relación al número de citas. En los primeros años WoS era algo superior (Bar-Ilan & Levene, 2007), aunque sólo un año después y para 11 revistas analizadas GA recuperaba 1.313 citas frente a 1.111 de WoS. El solapamiento entre ambas fue sólo del 56,93% (Kousha & Thelwall, 2008). Harzing (2013) también contabilizó un mejor rendimiento de GA. No se calculó la tasa de citas únicas ni de solapamiento.
- Química. El área de Ciencia donde GA ofrece peores resultados. Levine-Clark & Kraus (2007) comparan GA y Chemical Abstracts Service (CAS) a través de una serie de búsquedas. Las más fiables fueron la búsqueda por componente donde CAS recupera el 96,7% de los trabajos mientras que GA recupera sólo el 51,1%, y la búsqueda por autor, en la que CAS encuentra el 97,8% y GA un 57%. A pesar que desde el 2008 GA indexa la editorial ACS, en el último análisis de citas identificado WoS aún recuperaba un número mayor de citas para esta área (Harzing, 2013).
- Ecología. Christianson (2007) encontró para el área de Ecología que GA recuperaba el 88,6% de los 840 registros de la muestra, aunque con una baja calidad de los registros para el 14,5%.

- Ciencias de la Naturaleza. Respecto a 29 autores de la Universidad de Bergen, WoS recupera 1.573 registros mientras GA recuperó 5.048, mostrándose como una fuente más exhaustiva. La tasa de cobertura frente a WoS fue del 86%.

2.3.3.- Ciencias Sociales.

Una de las áreas más polémicas debido a la ausencia de trabajos “puros de cobertura” lo que obliga a orientarse a través de los análisis de citas. Éstos muestran en términos generales una recuperación de citas muy superior de GA sobre WoS y Scopus, aunque cuando estos trabajos se acompañan de un estudio de la tipología documental respecto a las obras citantes parece que GA no recupera una buena parte de las revistas nucleares (Bergman, 2012). En Norris & Oppenheim (2007), GA presenta un 56% de revistas indexadas consistentemente, mientras que la tasa de recuperación de artículos es muy similar a Scopus y superior al 95%. En Kousha et al. (2011) se encontró que GA perdía el 54,93% de las citas de Scopus para el área de derecho, así como el 31,47% para Sociología, el 52,27% para Arqueología y el 48,63% para Ciencias Políticas. Dos áreas dentro de las Ciencias Sociales que han recibido un especial tratamiento han sido:

- Economía. Una de las disciplinas de Ciencias Sociales mejor representada en el índice de GA. En Neuhaus et al. (2006) GA presenta una cobertura de tan sólo el 52% respecto a ABI/INFORM, pero esto se puede deber a que ABI contiene mucha prensa económica que GA no indexa. En Mingers & Lipitakis (2010) en cambio, GA recuperó el 66% de una muestra de 4.600 trabajos. Para los artículos de revista tuvo una cobertura muy buena del 89% mientras WoS sólo recuperaba el 47,6%. En los análisis de citas, GA suele recuperar un número de citas muy superior a WoS gracias a la indexación de bases de datos de acceso libre como RePEC donde aparecen otras tipologías documentales, destacando el trabajo de Harzing (2013) en el que quintuplica a WoS.
- Ciencias de la información y biblioteconomía. En Lewandowski (2010) GA recupera el 91,6% de una muestra de 3.580 artículos de 35 revistas internacionales de máximo nivel entre 2004-2006, lo que demuestra un alto grado de indexación para las revistas nucleares de ámbito internacional, mientras que en BA-Ilan (2010) GA recupera el 94,79% de las citas recuperadas por WoS y Scopus.

2.3.4.- Arte y Humanidades.

Para esta área de la ciencia prácticamente no se han realizado estudios de cobertura en GA y aún son menos los que han utilizado una metodología rigurosa. Hay que volver al estudio de Neuhaus et al. (2006) que a partir de una muestra aleatoria sobre 50 registros para cada base de datos encontró una tasa de cobertura en GA muy pobre, del 8% sobre Art Abstracts, 6% Historical Abstracts, 6% IIMP, 8% MLA Bibliography y un 22% sobre Philosopher's Index, con una media de cobertura para todo el área del 10%. Rozear (2009) realiza un interesante estudio sobre arte a partir de una muestra (no aleatoria) de 472 artículos de revista. GS presenta un 35% de cobertura, superado por Art Full Text (42%), Bibliography of the History of Art (54%) y por Arts & Humanities de WoS (73%). Kousha et al. (2011) encontró que a pesar de recuperar GA muchas más citas, para Historia la tasa de citas únicas de Scopus frente a GA era altísima, del 67,29% (citas en Scopus que no se encuentran en GA) y del 41,46% en Filosofía (en este estudio no se realizó una comprobación de búsqueda de los trabajos por el título). Todos estos estudios indican que GA no es capaz de recuperar consistentemente las revistas nucleares para el área de humanidades.

3.- ESTUDIO DE COBERTURA EN EL ÁREA DE ARQUITECTURA

3.1.- Metodología.

3.1.1.- Avery: base de datos referencias para el área de Arquitectura

Elaborada por "The Avery Architectural and Fine Arts Library" de la Universidad de Columbia, con cerca de 8000.000 registros es una de las bases de datos más importantes del mundo para esta área. Indexa alrededor de 700 revistas, en algunos casos desde finales del siglo XIX. También ofrece más de 30.000 obituarios. Las principales materias que cubre son la Arquitectura y Arquitectura del diseño, Diseño Interior, Planificación urbana, Historia de la Arquitectura, mobiliario y decoración, arqueología, Paisajismo, etc. (ProQuest, 2014). Las búsquedas por lengua realizadas a finales de Agosto (2014) ofrecen los siguientes resultados: inglés 498.889 registros, alemán 77.025, francés 55.524, italiano 52.624, español 34.924, japonés 21.293, etc. El CNEAI utiliza esta base de datos como fuente referencial en la disciplina para la evaluación académica, <http://boe.es/boe/dias/2013/11/21/pdfs/BOE-A-2013-12234.pdf>.

3.1.2- Periodos de estudio dentro de la base de datos Avery

Para identificar posibles sesgos en la cobertura según el año de publicación, se han seleccionado tres períodos de estudio independientes:

- 1950-1989. Época anterior al nacimiento de la web. En principio y debido a que GA se nutre de registros publicados online, la cobertura de GA sobre Avery para este período debería ser inferior.
- 1990-2009. Época web.
- 2010-2014. Con esta última muestra se intenta obtener una fotografía lo más actual posible del grado de cobertura de GA sobre Avery. Existen razones para creer que el nivel de indexación se haya podido incrementar, gracias por ejemplo a la consolidación del uso (aún moderado) por parte del personal académico de los repositorios institucionales o por el surgimiento en estos últimos años de las redes sociales científicas tipo Academia.edu o ResearchGate.

3.1.3.- Selección de la muestra.

Para la base de datos Avery se ha aplicado la primera de las técnicas utilizadas en Neuhaus et al. (2006), la cual requiere que cada registro de la base de datos tenga un número identificativo que permita su consulta. Este número en Avery se encuentra en de la etiqueta “número de acceso”. Posteriormente, se identificó que los valores para “el número de acceso” entre 63.987 y 295.120 hacían referencia mayormente a trabajos publicados entre 1950 y 1989, a la vez que para valores entre 295.120 y 654.500 se hacía referencia a trabajos publicados para el período 1990-2009, y los valores entre 654.501 en adelante se referían a publicaciones entre 2010 y junio de 2014. De esta forma se identificó las tres poblaciones del estudio.

Posteriormente, y a diferencia del trabajo de Neuhaus cuyas muestras fueron de tan solo 50 elementos, se ha tratado de dar al estudio un significado estadístico concreto. Desde la página web <http://es.gmi-mr.com/solutions/sample-size-calculator.php> se puede obtener automáticamente el tamaño de la muestra para una población de estudio dada y para un nivel de confianza y un intervalo de confianza prefijados. En este caso, se han tomado los valores 95% (nivel de confianza) y 5 (intervalo).

Posteriormente y para identificar los registros específicos que se debían consultar en Avery, se ha utilizado la página web Random.org, en concreto su aplicación “Random Integer Generator”, <http://www.random.org/integers/>, ya utilizada en los trabajos de

Neuhaus et al. (2006) y Chen (2010b), que permite obtener una serie de números aleatorios correspondientes al tamaño de la muestra y para el intervalo dado por los valores del “número de acceso” correspondientes a cada período.

Los datos cuantitativos para cada uno de los tres períodos se muestran en el Anexo I.

3.1.4.- Comprobación de registros en GA.

Una vez que los registros en Avery correspondientes a cada período fueron identificados, se realizó la búsqueda en GA individualmente para cada obra. Se han seguido los siguientes pasos:

- Se buscó por el título del artículo tanto mediante frase exacta (uso de comillas) como por palabra clave. Para aquellos trabajos con subtítulo, la búsqueda se realizó primero con el subtítulo y en caso de respuesta negativa sin el subtítulo.
- Si no se encontró el artículo, se utilizó la búsqueda avanzada para realizar búsquedas a través de la combinación de palabras significativas del título con el apellido del primer autor o con el título de la revista.
- Si tampoco se obtuvo un resultado positivo a través del paso anterior, en la búsqueda avanzada se llevó a cabo la consulta por el título de la revista y el apellido del primer autor, limitando aquella al año de publicación. De esta forma se localizaron publicaciones con pequeñas variaciones en el título. Para este último caso, a través del botón de GA “cita” se comprobó que coincidiese el número y paginación de la revista con el registro en Avery.

3.1.5.- Tipos de registros

Varios autores han enriquecido sus trabajos de cobertura con el estudio de los tipos de registros que la base de datos de GA contiene. Por ejemplo, Mayr P. & Walter A.-K., tomando como criterio la accesibilidad, diferencian registros con enlace al texto completo, registros con enlace a la referencia completa y normalmente al resumen, referencias bibliográficas sin ningún tipo de enlace y referencias de libros (2008). En el mismo sentido, pero aún más exhaustiva, Christianson M. clasificaba los registros de GA en 6 categorías: directos (enlace al texto completo libre), con barreras (enlace al texto completo bajo suscripción o pago), acceso al resumen, cita completa, cita parcial completada (cuando a través del enlace del registro se podía acceder a la referencia completa) y cita parcial no completada.

En este trabajo se pretende investigar también cómo la base de datos de GA es construida, por tanto no sólo interesa conocer el grado de accesibilidad de los registros, sino sobre todo el modo en que éstos son generados y las características que los determinan. De esta manera, se propone la siguiente clasificación parcialmente inspirada en las anteriores:

- Registros con acceso libre al texto completo.
- Referencias con enlace a un registro externo (bases de datos, página web del editor, redes sociales científicas, etc.)
- Referencias sin enlace.
- Citas.

Cada uno de los registros correspondientes a las tres muestras encontrado en GA se ha clasificado dentro de alguna de las anteriores categorías. También se ofrecen datos para las tres muestras de los registros con acceso al texto completo dentro del campus (UDC) gracias al Programa de Bibliotecas de GA.

3.1.6.- Fuentes de GA.

Al igual que en el apartado anterior, para todos los trabajos recuperados en GA se ha analizado las distintas fuentes (dominios) desde los que GA ha recuperado bien el texto completo o bien los metadatos utilizados para construir el registro. Estas fuentes se han clasificado de la siguiente manera:

- Página del editor (aquí se incluye tanto a grandes editores comerciales como a editoriales académicas al estilo de la Universidad Nacional Autónoma de México (UNAM, o las páginas web de las propias revistas)
- Bases de datos (en este apartado se incluyen agregadores comerciales tipo Ingenta, bases de datos especializadas, nacionales, etc. También se incluyen los repositorios temáticos).
- Páginas de universidad (páginas departamentales, lecturas recomendadas, etc.)
- Páginas personales (del propio autor)
- Repositorios institucionales.
- Redes sociales científicas (Academia.edu, ResearchGate, etc.)

Para los dos primeros puntos (página del editor y bases de datos) se presentan sendas tablas con los principales contribuyentes para cada período de estudio.

Hay que tener en cuenta que un trabajo puede haber sido indexado por parte de GA desde distintas fuentes, por lo que la consulta de éstas no se ha limitado a la fuente destacada por GA, sino que se ha realizado un estudio exhaustivo de éstas a través de la etiqueta “todas las versiones”. Para cada período se indica el número medio de fuentes por registro así como el nº de fuentes máximo para un solo registro. Un mayor número medio de fuentes por registro podría ser un indicador válido para determinar el grado de exhaustividad en la indexación por parte de GA de un área del conocimiento.

3.1.7.- Estudio por países.

En este caso sólo para el último período (2010-2014) se ha identificado para cada registro de la muestra el país de origen de la revista. No se ha tenido en cuenta el idioma, ya que al existir muchas publicaciones cuyos artículos se presentan en más de un idioma (normalmente lengua de origen e inglés), no parecía relevante estudiar la existencia de algún sesgo en el índice de GA por este criterio. Además, de esta forma, se ha podido calcular el grado de indexación, así como las correspondientes tipologías de registros y fuentes originales utilizadas para cada nación, pudiendo de alguna forma vislumbrar el mapa de la información científica nacional a la que GA tiene acceso por país.

3.1.8.- Búsqueda en Google.

También para el último período, se han buscado todos los trabajos de la muestra en el buscador general de Google. Con ello, se ha tratado de averiguar hasta qué grado las revistas de arquitectura ofrecen los contenidos de sus números online, qué estándares web utilizan y hasta qué punto GA falla a la hora de capturar esta información. La búsqueda se ha realizado por el título del artículo, bien como búsqueda exacta o como palabra clave. En algunos casos también se ha incluido en la caja de búsqueda, junto al título, el apellido del primer autor y/o el título de la revista. Se ha indicado para cada registro si Google recuperaba de alguna forma la referencia del artículo, aún de forma mínima (título y revista, título y autores, etc.), así como si ofrece el acceso al texto completo del artículo o al menos a la versión online de éste. También se indicaron las fuentes a través de la que se recuperó la información.

3.2.- Resultados.

3.2.1.- Resultados de cobertura para los tres períodos.

GA presentó una tasa de cobertura para los tres períodos de un 19,53%, 30,99% y 36,03% respectivamente como se muestra en la imagen siguiente.

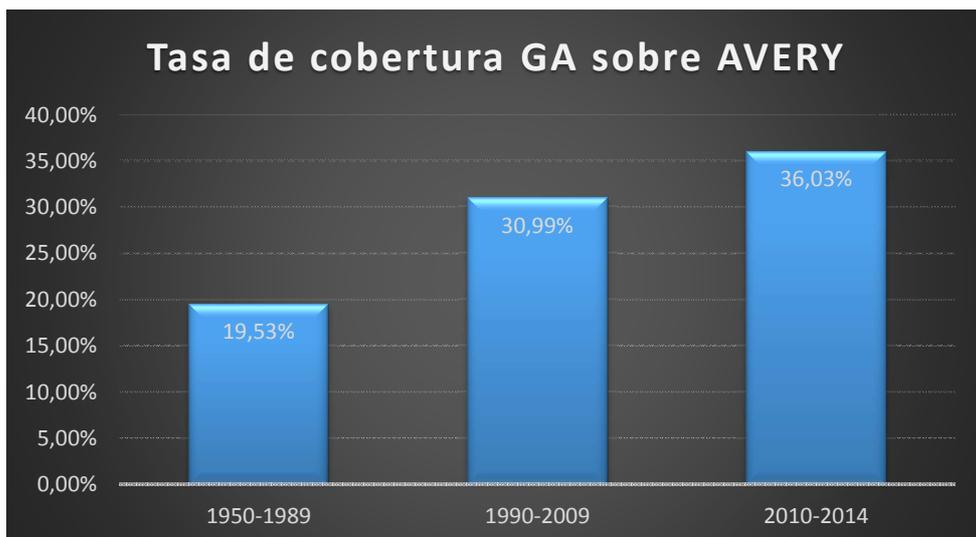


Gráfico 1: Tasas de cobertura para las muestras 1, 2 y 3.

De todas formas, para poner los datos anteriores en contexto, es necesario ofrecer una breve descripción tanto de la tipología documental de las publicaciones como de la calidad de los registros encontrados en Avery, ya que de alguna forma estas variables han podido incidir en algún grado en la tasa de recuperación de GA. Así por ejemplo, y como se puede observar en el diagrama de barras siguiente, en el primer período se encuentran 16 obituarios, mientras que para el segundo sólo se hallaron 2 y ninguno correspondiente al tercero. De los 18 obituarios GA sólo recuperó 1, mostrándose ineficaz para esta clase de publicaciones. En cambio, GA parece mucho más eficaz en recuperar revisiones de libros, una tipología documental que se incrementa ostensiblemente para los dos últimos períodos. Por otra parte, y en relación a la calidad de los registros, 171 trabajos aparecen en Avery sin mención de autoría para el primer período, mientras que para los dos períodos siguientes se reducen a 100 y a 70. De todo lo anterior se abundará en el apartado de discusión.

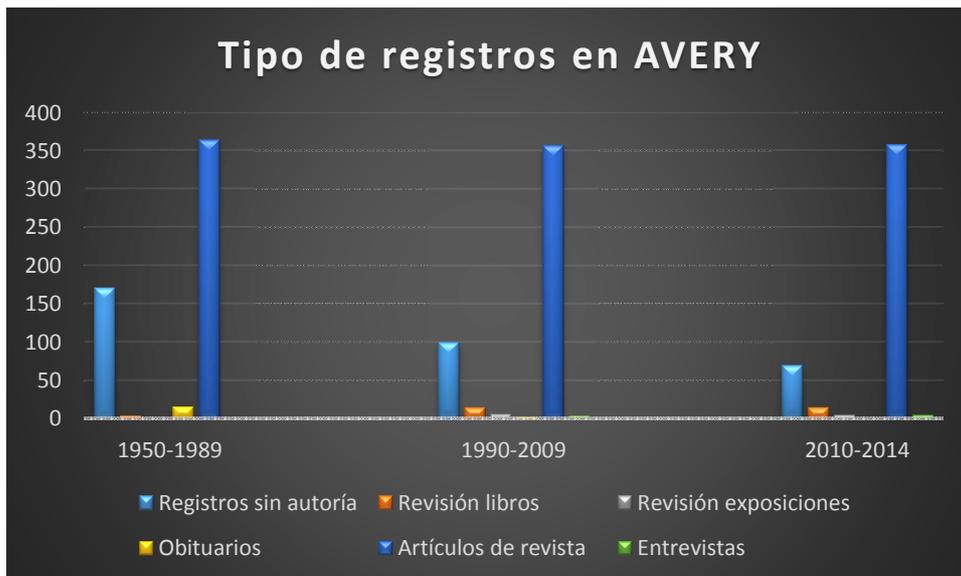


Gráfico 2: tipos de registros en la base de datos Avery

3.2.2.- Tipo de registros y fuentes.

Un estudio de la tipología de los registros de GA puede dar una idea de cómo el buscador construye su índice para el área de Arquitectura, de qué modo es capaz de rastrear los trabajos publicados y de qué forma se modifica ese patrón según el año de publicación a lo largo de más de 6 décadas.

3.2.2.1.- Período 1950-1989

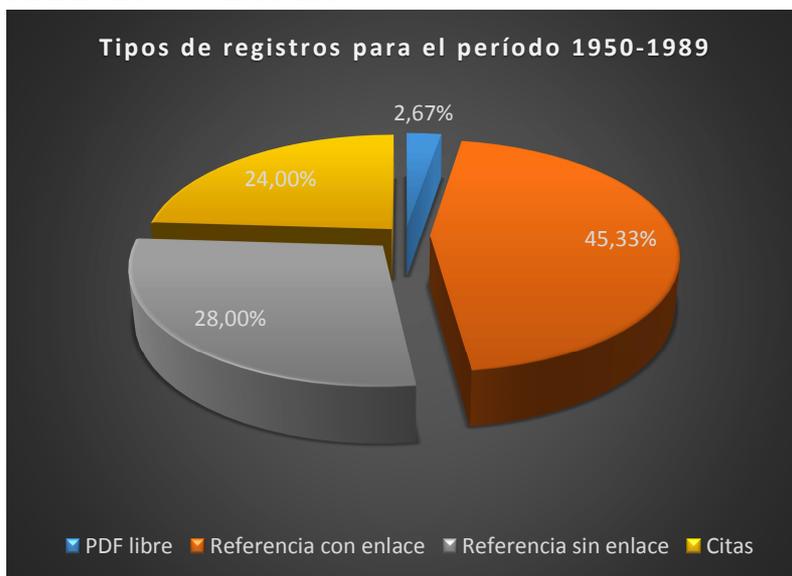


Gráfico 3: Tipos de registros en GA, muestra 1

Para el primer período (imagen superior) se observa que una parte muy importante de los registros (52%) no presentan enlace. Esto es una limitación importante para los

usuarios que no tienen la posibilidad de consultar el texto completo del trabajo, el resumen del mismo o al menos una referencia completa externa. En concreto, el 24% de los registros son recuperados de forma indirecta mediante cita, mientras que el 28% restante corresponde a referencias sin enlace cuya fuente es desconocida para el usuario. Esta última tipología es especialmente preocupante tanto porque no permite averiguar el origen de los metadatos rastreados por GA (por lo que no se puede verificar la autoridad de la fuente) como por la baja calidad bibliográfica que la misma ha mostrado a lo largo de la presente investigación.

Por otro lado, sólo 36 registros (48% del total) presentan una referencia bibliográfica con enlace externo. En este caso sí es posible realizar un estudio sobre las páginas web y bases de datos de las que GA se ha alimentado. En total, GA ha utilizado 48 enlaces externos y 22 fuentes únicas, lo que resulta una media de 1,33 enlaces externos por registro, un número bajo que pone de relieve la poca incidencia de la Arquitectura en los recursos web disponibles por GA.

Fuentes (1950-1989)	
Registros con enlace	36
Total enlaces externos	48
Fuentes únicas	22
Enlaces por registro	1,33

Tabla 2: Registros con enlaces externos y fuentes únicas para la muestra 1

Tipología	Total enlaces	Porcentaje enlaces	Fuentes únicas
Editores	16	33,33%	10
Bases de datos	32	66,66%	12

Tabla 3: Tipología de las fuentes de GA para la muestra 1

Como se puede observar en las tabla 2, 3 y 4 sólo 16 de los 384 registros de la muestra 1 han sido recuperados por GA a través del editor de la revista. Taylor & Francis es la editorial más productiva con 5 enlaces recuperados. Pero estos 5 enlaces sólo hacen referencia realmente a 2 revistas: "American Institute of Planners Journal" (4 enlaces) y "Journal of Architectural Education" (1 enlace). Lo mismo sucede con Liverpool University Press, cuyos 2 enlaces se dirigen a sendos artículos de la revista "Town Planning Review". En total, de más de un centenar de revistas de arquitectura de la muestra 1, sólo 12 han sido recuperadas por GA total o parcialmente a través de la página del editor.

Editor	Frecuencia	BDs	Frecuencia
Taylor & Francis	5	JSTOR	12
Liverpool U. Press	2	Dialnet	5
Cambridge U. Press	2	bcin.ca	4
Sage	1	CAT.INIST	3
Wiley	1	OSTI	1
Elservier	1	RePec	1
Springer	1	TRID	1
Bretschneider	1	lw20	1
Pág. Revista "Antiquity"	1	Compludoc	1
"Middle East Technical U."	1	SPARA & BEVARA	1
		RAND	1
		Érudit	1

Tabla 4: Fuentes únicas de GA para la muestra 1

Otra cuestión de interés puede ser averiguar cuántos artículos de la muestra 1 ha podido GA indexar completamente. Este proceso de "escaneo" facilita una de las principales virtudes del buscador que es la exhaustiva recuperación de los trabajos desde su base de datos gracias a la búsqueda a texto completo. De los 10 editores de la tabla superior, 8 son grandes editoriales académicas con las que GA tiene firmado en su mayoría algún tipo de acuerdo. Siendo optimistas, GA podría indexar a texto completo un máximo de 14 trabajos por esta vía. Si se suman los 2 trabajos de la muestra que ofrecen un acceso libre al texto completo, 16 trabajos de los 75 recuperados podrían ser indexados al texto completo, es decir, un 21,33%, que descendería drásticamente al 4,16% si se tiene en cuenta el total de la muestra.

En relación a las bases de datos, GA ofrece 32 enlaces externos procedentes de 12 fuentes únicas. Es patente el fuerte desequilibrio entre la aportación de cada una de ellas. Así por ejemplo, las dos primeras fuentes recuperan más del 50% de los enlaces. Es de destacar en este sentido los 12 enlaces que recupera GA a través de JSTOR, una base de datos de carácter retrospectivo y con un abundante contenido en ciencias sociales, artes y humanidades.

Es importante señalar la ausencia de bases de datos específicas de Arquitectura. Salvo bcin.ca (4 enlaces), base de datos canadiense sobre conservación y restauración de objetos culturales, y SPARA & BEVARA (1 enlace), relacionada con la eficiencia energética en edificios históricos, el resto de bases de datos son multidisciplinares (destaca la aportación de la española Dialnet con 5 enlaces y la francesa CAT.INIST con

3 enlaces) o incluso de áreas aparentemente ajenas a la Arquitectura como por ejemplo RePec (economía), TRID (transporte) o OSTI (energía).

3.2.2.2.- Período 1990-2009.



Gráfico 4: Tipos de registros en GA para la muestra 2

De los 119 trabajos recuperados por GA para la muestra 2, más de la mitad, 60 (50,42%), son registros sin enlace de lo que poco o nada se sabe sobre su fuente de origen. Otros 13 trabajos (10,92%) son recuperados de forma indirecta mediante cita. Por tanto, el 61,34% de los trabajos recuperados para este período no ofrecen ningún tipo de enlace externo, con una información bibliográfica limitada o pobre. GA es capaz de recuperar 46 registros con enlaces externos, de los cuales 8 (6,72%) permiten el acceso libre al texto completo.

Fuentes (1990-2009)	
Registros con enlace	46
Total enlaces externos	58
Fuentes únicas	24
Enlaces por registro	1,26

Tabla 5: Registros con enlaces externos y fuentes únicas para la muestra 2

En relación al número de enlaces externos y fuentes únicas, este período muestra un patrón similar al anterior aunque con algunas pequeñas diferencias reseñables. El número medio de enlaces por fuente sigue siendo muy bajo, de 1,26, aún inferior al correspondiente en la muestra 1, indicador que señala de otra forma a la tasa de

cobertura la debilidad del índice de GA para el área de Arquitectura. De hecho, los tres artículos que presentan un mayor número de enlaces (3 cada uno) y de citas, son propios más del área de Urbanismo que de la Arquitectura.

Título	Año	Revista	Nº citas	Nº enlaces
Compact, diffuse, or would-be discipline? Assessing cohesion in planning scholarship, 1963-2002	2006	<u>Journal of planning education & research</u>	13	3
The cycle of fragmentation and sprawl: a conceptual framework and empirical model	2006	<u>Environment & planning B, planning & design</u>	39	3
Trajectories of neighborhood change: the case of gentrification	1990	<u>Environment & planning</u>	124	3

Tabla 6: Trabajos con el mayor número de enlaces externos para la muestra 2

Por otra parte, la muestra 2 ofrece un mayor número de fuentes únicas así como una ampliación en su tipología. Con la aparición de la tecnología web aparecen nuevas posibilidades de difusión de la ciencia. Así, 4 artículos son accesibles online a través de repositorios institucionales, mientras 2 artículos más fueron recuperados a través de las webs de sendas universidades. En cualquier caso, el mayor aporte de registros procede de nuevo de las bases de datos con el 60,34% de los enlaces externos. En cambio, las páginas de los editores bajan tanto en la aportación porcentual de enlaces externos, que desciende hasta el 29,31% como en el número de fuentes únicas, pasando de 10 a 6.

Tipología	Total enlaces	% enlaces	Fuentes únicas
Editores	17	29,31%	6
Bases de datos	35	60,34%	12
Repositorio institucional	4	6,89%	4
Páginas universidad	2	3,44%	2

Tabla 7: Tipología de las fuentes de GA para la muestra 2

En la tabla siguiente se puede apreciar cómo la editorial Taylor & Francis casi duplica a las otras 5 editoriales, siendo dos de ellas específicas a una revista concreta. De este modo, a través del acuerdo con las editoriales, GA podría tener acceso al texto completo de tan solo 14 trabajos en el mejor de los casos, un 3,64% del total de la muestra.

Editor	Frecuencia	BDs	Frecuencia
Taylor & Francis	11	Dialnet	11
E&V	2	elibrary.ru	7
Oxford J.	1	Persee.fr	3
Sage	1	RePEC	2
Wiley	1	Ingenta	2
Landscape J.	1	bcin.ca	2
		Informit	2
		RACO	1
		CAT.INIST	1
		Compludoc	1
		TRID	1
		Safetylit	1
		Finehomebuilding	1

Tabla 8: Fuentes únicas en GA, muestra 2

Respecto a las bases de datos, ha desaparecido JSTOR, para tomar su relevo proyectos de carácter nacional y multidisciplinar como Dialnet (España, 11 enlaces), elibrary.ru (Rumanía, 7 enlaces), Persée (Francia, 3 enlaces) o Informit (Australia, 2 enlaces); así como bases de datos especializadas en áreas ajenas a la Arquitectura, como RePEC (economía, 2 enlaces), TRID (transporte, 1 enlace), Safetylit (medicina, 1 enlace), etc. Sólo Bcin.ca con 2 enlaces y Finehomebuilding con 1 enlace se pueden considerar bases de datos especializadas directamente relacionadas con la Arquitectura.

3.2.2.3.- Tercer período. 2010-2014.



Gráfico 5: Tipos de registros en GA para la muestra 3

En el tercer período de estudio se ha encontrado un cambio sustancial en la composición de los registros según su tipología. Así, ahora son los registros con enlace externo los predominantes alcanzando la suma de las dos categorías (con link a una referencia externa o al texto completo) el 65,33% del total de los trabajos recuperados. Por el contrario, tanto los registros sin enlace externo como las citas recuperadas bajan ostensiblemente respecto al período anterior (del 50,42% al 32,61% y del 10,92% al 2,17% respectivamente). De lo anterior, parece que se pueden destacar dos ideas:

- GA no sólo es capaz de recuperar una mayor cantidad de trabajos para los últimos 4 años y medio que para los dos períodos anteriores, sino que también los registros que ofrece son de mayor calidad bibliográfica y con un mayor acceso a la información primaria.
- Además, la reducida aportación a través de cita (indexación indirecta) hace pensar que el índice de GA aún tiene un importante margen de mejora en la cobertura para este período.

El estudio de los enlaces externos y las fuentes únicas permite un análisis más detallado de estos resultados. En la tabla 8 se observa que además de un importante incremento en el número de registros con enlace, pasando de 46 a 90 (un incremento porcentual del 95,65%) respecto al período anterior, para el caso del total de enlaces externos el crecimiento aún es mayor pasando de 58 a 145 (incremento porcentual del 150%). El número medio de enlaces por registro aumenta hasta un 1,61, una cifra que aún puede considerarse baja. La moda fue 1.

Fuentes (2010-2014)	
Registros con enlace	90
Total enlaces externos	145
Fuentes únicas	57
Enlaces por registro	1,61

Tabla 8: Registros externos y fuentes únicas, muestra 3

Los artículos con más enlaces en GA vuelven a corresponder a revistas más cercanas al urbanismo que a la Arquitectura propiamente dicha, aunque como es obvio ambas áreas del conocimiento están fuertemente entrelazadas (tabla siguiente). El artículo más citado, con 25 citas y 3 enlaces externos, también corresponde a la revista "Environment & Planning". Para las tres muestras se aprecia una fuerte correlación entre el número de citas y el número de enlaces externos.

Título	Año	Revista	Nº citas	Nº enlaces
Can community design build trust? A comparative study of design factors in Boise, Idaho neighborhoods	2010	Cities	17	4
Proximity and perceived safety as determinants of urban trail use: findings from a three-city study	2010	Environment & planning	6	4

Tabla 9: revistas con mayor número de enlaces externos, muestra 3

El fuerte incremento del total de enlaces externos respecto al período anterior (que como se ha visto fue de un 150%) no ha venido provocado por una mayor productividad en las fuentes únicas, sino más bien por un fuerte incremento en el número de éstas así como en la aparición de nuevas modalidades. Aunque las bases de datos con un 56,25% del total de enlaces externos siguen siendo el recurso básico de GA para el área de Arquitectura, para los últimos 4 años se ha producido un fuerte incremento en el uso de otros recursos como los repositorios institucionales. Así, mientras que para la muestra 1 no se recupera ningún trabajo de este modo y para la muestra 2 se recuperan 4 (6,89%), la cifra se incrementa hasta los 16 trabajos recuperados para el período 2010-2014 (incremento porcentual del 300%), representando el 11,03% del total de los enlaces externos. Además, surgen una nueva tipología de fuentes, las redes sociales científicas, a través de las que GA recupera 3 enlaces (2,06%), 2 desde la red social ResearchGate y 1 desde Academia.edu. Para el caso de ResearchGate, se permite el acceso libre al texto completo a través de la plataforma.

Tipología	Total enlaces	% enlaces	Fuentes únicas
Editores	40	27,58%	14
Bases de datos	81	55,86%	21
Repositorio institucional	16	11,03%	16
Pág. Universidad	3	2,06%	3
Redes sociales	3	2,06%	2
Pág. Personales	2	1,37%	2

Tabla 10: Tipos de fuentes de GA muestra 3

También destaca la indexación de GA de pequeñas editoriales de carácter académico, sin excesiva repercusión internacional (como sí era el caso de Liverpool University Press y sobre todo Cambridge University Press para el período 1) y con una procedencia geográfica muy determinada (iberoamericana). Es el caso de las editoriales de la Universidad Politécnica de Cataluña, la Complutense (esta con 4 enlaces), la Universidad de Chile y la Universidad Autónoma de México. También es de reseñar el incremento en el número de revistas que GA indexa a través de su propia página de la publicación. A la ya conocida “Environment & Planning”, se suman “Journal of Green Building”, “Bitácora Urbano/Territorial”, “Urbani Izziv” y “Journal of Sustainable Real Estate”. En total se trata de 14 fuentes únicas frente a las 6 del período anterior.

Editor	Frecuencia	BDs	Frecuencia
Taylor & Francis	14	Dialnet	29
Wiley	4	Ingenta	6
Springer	3	RePEC	5
Env. & Plann. (revista)	6	CNKI	5
Editorial UCM (académica)	4	Redalyc	4
Editorial UPC (académica)	1	cat.inist.fr	4
Editorial Univ. Chile (académica)	1	NAUKI	4
Editorial UNAM (académica)	1	CORE	3
Journal of green building (revista)	1	JSTOR	3
Bitácora urbano/Territorial (revista)	1	RACO	3
Urbani Izziv (revista)	1	Scielo	3
Journal of sustainable real estate (revista)	1	Informit	2
American Real Estate Society (editor)	1	CEEOL	2
Elsevier	1	CSIC	1
		TRID	1
		cqvip.com	1
		PubMed (EEUU)	1
		Europa PubMed Central	1
		Maneyonline	1
		Wanfangdata	1
		MUSE	1

Tabla 11: Fuentes únicas para la muestra 3

Respecto a las bases de datos se observa un fuerte incremento en el número de fuentes únicas, pasando de 12 a 21 en relación al período anterior. Se identifican nuevas bases de datos de carácter nacional (o regional) y multidisciplinarias que se vienen a unir a las ya vistas Dialnet (esta vez con una sorprendente cifra de 29 enlaces externos), CAT.INIST o Informit. Algunas de ellas son las españolas RACO y CSIC, las latinoamericanas Redalyc y Scielo, de Europa del Este CEEOL y NAUKI (Polonia), procedentes de China Wanfangdata, cqvip.com y CNKAI, etc. Persisten bases de datos especializadas de otras áreas como PubMed (salud), RePEC (economía) o TRID (transporte). Como se puede ver en la tabla anterior, no se encontró ninguna base de datos específica de Arquitectura.

3.2.3.- Estudio por países (período 2010-2014).

Al analizar los datos ya vistos para la muestra 3 poniendo el foco en los países de origen de las revistas, se identifican una serie de pautas nacionales que permiten entender en mayor grado cómo se construye el índice de GA para el área de Arquitectura, además de contrastar la existencia de distintos modelos de difusión científica.

Los 382 registros de la muestra corresponden con 139 títulos de revista, los cuales se distribuyen geográficamente en 31 países. Los datos de cobertura, así como la tipología de registros y enlaces externos son presentados para cada uno de ellos (tabla Anexo III).

Inglaterra y EEUU son los países con una mayor participación en la muestra, con 89 y 92 trabajos cada uno. Ambos ofrecen dos patrones de información científica bien diferenciados. Inglaterra, con un 42% de cobertura de GA sobre sus trabajos, presenta un amplio abanico de fuentes. Entre ellas destacan las páginas del editor. De los 37 trabajos recuperados por GA, 18 (48,64%) han sido recuperados desde la página web del editor o de la propia revista, lo que debería garantizar una calidad bibliográfica alta. Otros 14 artículos son recuperados gracias a bases de datos de acceso libre. Además, los autores de los trabajos han mostrado una fuerte inclinación a incrementar la visibilidad de sus obras, recuperando GA 7 trabajos desde repositorios institucionales, 4 de páginas web y páginas de autor y 2 desde redes sociales científicas. Por el contrario, EEUU con un 30% de tasa de cobertura en GA (28 trabajos), sólo presenta 10 registros con enlace, mientras que los otros 18 son registros sin enlace de origen indeterminado. Tan sólo recupera 2 trabajos desde repositorios institucionales.

España, China y Australia presentan una alta tasa de cobertura debido principalmente a poseer bases de datos nacionales de acceso libre cuyo objeto es registrar y difundir la producción científica del país. De todas formas existen algunas características diferenciadoras.

España presenta de las tres la tasa de cobertura más baja con un 56%, pero esto es debido principalmente a la incompleta indexación que GA realiza de Dialnet. Los 15 registros que recupera tienen enlace externo, 5 de ellos al texto completo. Además, GA recupera 5 artículos desde la página web del editor (todas ellas son editoriales universitarias), 14 a través de la base de datos Dialnet y para 4 de ellos se facilita el texto completo a través del uso de repositorios institucionales. Todos ellos son pues de una calidad bibliográfica alta.

Para el caso de China, GA es capaz de recuperar el 70% de los trabajos a través de bases de datos de su país (CNKI y Wanfangdata) cuyo objeto es tratar de dar una difusión internacional de la producción científica propia (en la base de datos CNKI se utiliza también el idioma inglés).

Australia sea quizá el modelo a seguir en relación a la difusión internacional de sus publicaciones. Sus revistas han logrado ser publicadas por editoriales de difusión internacional, como Taylor & Francis; tiene una base de datos nacional, Informit; y sus autores difunden los trabajos a través de los repositorios institucionales y redes sociales científicas. A la postre, la cobertura de GA a sus trabajos es del 100%. De los 8 trabajos recuperados, todos ellos son registros con enlace, 7 de ellos a través del editor, y tres de ellos con acceso libre al texto completo. Todos los registros son pues de calidad bibliográfica alta.

A diferencia de los tres países anteriores, Alemania, Italia y Holanda ofrecen un panorama desolador. Para el caso alemán, GA sólo recupera el 22% de los trabajos (9 de 41). De ellos, sólo 3 presentan algún tipo de enlace, bien a través de la página del editor (1) o de bases de datos (3), mientras que los otros 6 son recuperados por GA a través de registros sin enlace y por tanto de procedencia difícil de determinar. No se ha rastreado ningún trabajo desde repositorios institucionales, páginas personales o redes sociales científicas.

Italia, con un 42% de cobertura sobre sus artículos, podría considerarse un país intermedio en el grado de representación en GA, pero un análisis detallado ofrece una imagen diferente. De los 8 trabajos que recupera, 4 son obtenidos gracias a la base de datos española Dialnet, gracias a una política un poco extraña por parte de los

administradores que no permiten indexar artículos sueltos de autores españoles en publicaciones extranjeras. Realmente, sólo los dos artículos de la revista “Nexus Network Journal” ofrecen una buena visibilidad a través de la editorial Springer y bases de datos internacionales.

Por último para Holanda, un país con gran tradición editorial, GA sólo recupera 1 trabajo de los 7 que aparecen en la muestra. Además, el registro carece de enlace externo alguno.

Visto los países analizados anteriormente, se puede concluir que para que un país se favorezca de una visibilidad importante dentro del índice de GA se requiere la combinación en cierto grado de distintos factores: presencia de sus revistas en editoriales internacionales o al menos con una infraestructura digital específica, bases de datos nacionales, así como una actividad de difusión por parte de los autores.

3.2.4.- Búsqueda en Google

Todos los trabajos de la muestra 3 fueron de nuevo consultados, esta vez en el buscador general de Google. Los resultados indican que Google es capaz de mostrar algún tipo de registro sobre 296 de los 382 trabajos de la muestra (tasa de recuperación del 77,49%), a la vez que dirige al usuario al texto completo o a la versión online de 101 documentos (26,44%). En cambio, GA sólo recupera para ese período 138 (36,13%) y ofrece enlace al texto completo de 55 (14,39%). Las posibles explicaciones a estas diferencias son tratadas en el apartado de discusión.

3.3.- Discusión

3.3.1.- Editoriales STM y bases de datos.

Un aspecto que requiere comentario es el pequeño número de registros recuperados desde los portales de los editores, cuando el auténtico germen de GA se encuentra en la firma (secreta) de acuerdos con las principales editoriales académicas (Notess G.R., 2005). En cambio, para las tres muestras los registros recuperados por esta vía han sido pocos: 16 (4,16%), 17 (4,42%) y 40 (10,47%) respectivamente.

El problema parece radicar, como ya se apuntó en la primera parte del trabajo, en el carácter STM de la mayor parte de aquellas editoriales en las que áreas “limítrofes de la ciencia” como la Arquitectura del diseño no están apenas representadas. Así, y determinado por las propias características de la disciplina, el área de Arquitectura parece carecer de un núcleo duro de revistas de ámbito internacional. Para Ann

Forsyth, Arquitectura y “Ciencia” se desarrollan en diferentes espacios ajenos el uno al otro y donde el profesor de Arquitectura no se encuentra especialmente interesado en publicar según el concepto tradicional de la ciencia (Forsyth, 2007). De hecho, algunos autores consideran incluso polémica la expresión “investigación en Arquitectura del diseño” (2011), de ahí que la mayor parte de las revistas no se ajusten a criterios científicos. Ángela Sorli y Gonzalo Mochón, en un estudio sobre la visibilidad e internacionalización de las revistas españolas de Arquitectura, destacan la ausencia de la simple mención de autoría en muchos artículos (2012), mientras que para Javier Monedero buena parte de las revistas de arquitectura se limitan “únicamente a mostrar fotografías e informes descriptivos sobre aspectos relevantes de un proyecto” (Viñas A., Minobis E. & Millares R., 2007).

Lo mismo parece suceder con las bases de datos. Aunque han sido el tipo de fuente más productivo con 32, 35 y 81 enlaces desde GA, su comportamiento ha sido errático, en el sentido que no se ha identificado una base de datos propia del área que fuese capaz de surtir eficazmente al buscador científico. Aquí hay que volver de nuevo a lo comentado en la primera parte del trabajo. Consultando los repositorios en OpenDoar (Directory of Open Access Repositories), <http://www.opendoar.org/>, por el tamaño de sus bases de datos no se encuentra entre las primeras posiciones ninguna base de datos específica en Arquitectura.

3.3.2.- Indexación de pequeñas editoriales por GA.

El período 2010-2014 presenta un incremento sensible de trabajos recuperados por GA desde páginas de editores y páginas de revistas, pasando de 6 fuentes únicas a 14 respecto a la muestra anterior. Este incremento no se produce gracias a las grandes editoriales, sino precisamente a la indexación de editoriales universitarias de tamaño medio (UCM, UPC, UNAM y Universidad de Chile) y de páginas de revistas concretas de países con perfiles muy diversos: “Journal of Green Building” y “Journal of Sustainable Real Estate” (ambas de EEUU), “Environment & Planning” (Gran Bretaña), “Urbani Izziv” (Eslovenia) y “Bitácora Urbano/Territorial” (Colombia).

Esto es posible ya que cualquier revista con contenidos académicos o científicos puede solicitar su inclusión en el índice de GA cubriendo una solicitud web, <https://support.google.com/scholar/troubleshooter/2898950?rd=1>, y cumpliendo una serie de requisitos técnicos muy similares a los comentados por Arlitsch & O’Brien (2012) y que se detallan en la página de ayuda del propio buscador científico: publicación de los sumarios o del conjunto de los artículos de forma sencilla (por

ejemplo, ordenados cronológicamente), cada artículo o resumen del mismo debe tener su propia URL, no debe encontrarse a más de 10 links de la página principal, formatos HTML o PDF, peso del archivo no superior a los 5MGs, uso de esquemas de metadatos académicos (recomiendan varios como Highwire Press, Eprints, BE Press, etc.), acceso a las referencias bibliográficas de los trabajos, etc. Se trata realmente de requisitos fácilmente aplicables para un programador.

Por otro lado, existen softwares específicos para la gestión de una revista académica, muchos de los cuales facilitan la indexación de GA. Así por ejemplo, el software libre OJS (probablemente el más conocido), <https://pkp.sfu.ca/ojs/>, permite la indexación por parte de GA de forma casi inmediata, como indica la propia página de ayuda de GA para webmasters, <http://scholar.google.com/intl/es/scholar/inclusion.html>. El uso de servicios de hosting especializados como MetaPress o Highwire Press también permite este fin. Finalmente, para el caso de las editoriales universitarias, les basta con utilizar la infraestructura del repositorio de su institución.

Para responder a la pregunta de por qué las revistas y pequeñas editoriales de Arquitectura no solicitan la admisión en GA y configuran su sitio web al efecto, se analizó la página web de las 50 primeras revistas incluidas en la muestra 3 (2010-2014). En general, Las páginas de las revistas de Arquitectura así consultadas se han construido a través de gestores de contenidos estándar, primando la imagen y el diseño por encima de otras variables. En cambio, los sumarios no tienen un papel protagonista en absoluto. De hecho, 18 de ellas (36%), principalmente magazines, no ofrecían acceso a ningún sumario, lo que denota la diferencia del área de Arquitectura con la difusión online habitual de las revistas científicas puesta de manifiesto por Chen (2010a).

La mayor parte de estas revistas presentaban además referencias muy pobres, sin la estructura web mínima exigida por GA. Así, a veces se reducían a los títulos de los artículos ("Architectural Records"), en otros casos se ofrecía directamente el acceso a parte de la revista, normalmente las primeras páginas, a través de servicios de publicación online tipo ISSUU ("Architecture Minnesota" y otros), mientras que en algún caso publicaban todos los sumarios de la revista en un único archivo PDF ("Garden History"). En total, la media de años con sumarios accesibles online por revista fue de 13,1, destacando en este aspecto las pocas revistas en manos de grandes

editoriales. Por ejemplo, la revista “Journal of the American Planning Association” de Taylor & Francis ofrece el acceso a los sumarios por un período de 80 años.

Como conclusión a este apartado, se puede decir que el eminente carácter profesional y no científico de buena parte de las publicaciones hace que el cumplimiento de estándares académicos y la inclusión en buscadores científicos como GA no sea una prioridad.

3.3.3.- Evaluación científica. El movimiento Open Access (OA) y las redes sociales científicas.

La puesta en marcha de agencias evaluadoras de la investigación en la mayor parte de los países (en España las principales son la ANECA y el CNEAI) así como políticas de distribución de fondos basadas casi exclusivamente en la productividad científica, ha obligado al investigador a familiarizarse con conceptos como el factor de impacto o el “peer review”. Además, los nuevos procesos de evaluación que se están a implantar, como en el caso de Inglaterra con REF (Research Excellence Framework), apuestan de forma decidida por la flexibilidad y movilidad del investigador; por un perfil con capacidad para manejarse tanto en el mundo de la investigación, como en el de la empresa y la gestión pública; empujando al científico a la colaboración, tanto fuera de su institución como de su propia disciplina y donde cada vez cobra más importancia las tecnologías en red (Sumner, 2012; Brewerton, 2012).

En este nuevo contexto, tanto la visibilidad del investigador como de su producción científica se hace indispensable, independientemente que éste forme parte de una disciplina tradicionalmente ajena al modo científico.

El contexto anterior promete un incremento importante en la cobertura de GA para el área de Arquitectura a través de al menos tres caminos diferenciados:

- El movimiento de acceso abierto (OA).
- La creación de nuevas revistas de Arquitectura que se ajusten a los criterios de calidad de las agencias evaluadoras.
- Las redes sociales científicas.

El movimiento OA queda reflejado en la muestra 3 gracias a los 16 trabajos recuperados a través de repositorios institucionales y de las cuatro editoriales académicas (UCM, UPC, UNAM y Universidad de Chile) que ofrecen sus artículos a texto completo. En este sentido, hay razones para el optimismo, debido al incremento

de los mandatos de publicación en abierto, tanto a nivel institucional como a nivel nacional y europeo (para aquellas investigaciones apoyadas con fondos públicos), de la aprobación de normas como el Real Decreto 99/2011 (que obliga a colgar en abierto las tesis doctorales), así como a un incremento en el número de revistas científicas que permiten algún tipo de auto-archivo, siendo actualmente en España del 61% (Abad-García et al., 2012).

Por otro lado, las necesidades de los investigadores antes comentadas están creando una demanda creciente de revistas de Arquitectura que cumplan con los criterios de calidad exigidos por las agencias evaluadoras. En este sentido, nuevas revistas están tratando de cubrir el vacío existente. Un buen ejemplo de esto es la revista española “Progreso, Proyecto, Arquitectura”, <https://ojs.publius.us.es/ojs/index.php/PPA/>, surgida en 2010. Se cuenta el proceso de creación y cumplimiento de los criterios de calidad en un interesante trabajo (Ramos Carranza & Añón Abajas, 2012).

Por último, los propios investigadores necesitan maximizar su visibilidad como miembros de la comunidad científica, tanto para acceder a redes de colaboración, promoverse profesionalmente o simplemente mejorar la difusión de sus publicaciones. Los blogs, Facebook o Twitter han cumplido (siguen haciéndolo) una importante función de difusión y visibilización (Torres-Salinas & López-Delgado-Cozar, 2009), pero en los últimos años han surgido con gran fuerza las redes sociales científicas.

Quizás las redes sociales científicas más famosas sean ResearchGate, Academia.edu y Mendeley, aunque esta última, ahora en manos de Elsevier, ya no permite a los motores de búsqueda el acceso a sus bases de datos (Dobbs, 2013). Consultando el “Ranking Web de Repositorios” desarrollado por el CSIC, en el apartado “Top Portales”, http://repositories.webometrics.info/es/top_portales, ResearchGate y Academia.edu encabezan dicho ranking, ocupando la posición 3 y 5 por el tamaño de sus bases de datos. Además, Academia.edu encabeza el listado por el criterio de visibilidad y ResearchGate por la riqueza de sus registros. Ésta última red, con más de 3 millones de científicos registrados, permite el acceso libre a 11 millones de publicaciones y a 50 millones de abstracts según el magazine Deutschland.de (19 de Enero de 2014). Bien es cierto que su uso mayoritariamente se centra en las áreas “duras” (biología, medicina, informática, ingeniería, química, física y ciencias sociales). En una entrevista concedida a la revista IEEE Spectrum, Ijad Madisch, uno de los creadores de ResearchGate, afirmaba que aunque todas las disciplinas estaban

representadas en la red, “entre el 30 y 40% de los científicos pertenecían al área de la medicina y biología” (Cherry, 2013). GA indexa total o parcialmente Academia.edu y ResearchGate. Entre las dos se han recuperado sólo 3 registros para la muestra 3, un dato por el momento poco espectacular.

En este apartado se ha analizado el cambio que el mundo académico está sufriendo en dirección a una competitividad extrema y global y a una exigencia permanente de excelencia, presionando a los investigadores a una mayor productividad científica, a la elección de revistas que cumplan los criterios de calidad impuestos por las agencias evaluadoras y a la necesidad de mejorar la visibilidad personal y la de su producción científica. Todo ello está repercutiendo en un mayor acceso de los motores de búsqueda a los resultados de la investigación, bien a través del uso de los repositorios, de revistas OA o con estándares web académicos, así de las redes sociales científicas. Aunque la repercusión en la muestra 3 de estas fuentes aún es pequeña, se puede esperar un incremento importante en los próximos años ya que los investigadores de áreas con poca tradición científica como la Arquitectura probablemente tiendan a incorporar estas prácticas con cierto retraso en relación a otras disciplinas.

3.3.4.- Análisis de los registros de GA sin enlace externo ni citas recibidas.

Mucho se ha comentado del secretismo de GA sobre sus contenidos así como sobre las revistas que indexa. Algunos autores consideran que esto refleja no solo falta de transparencia sino también la incapacidad del buscador de controlar la calidad y autoridad de sus contenidos. Por ejemplo, GA ha reconocido que debido a que indexa artículos de un gran número de webs, no siempre puede afirmar dónde, o incluso si, un artículo ha sido realmente publicado (Gray J.E. et al., 2012), lo que provoca el escepticismo y rechazo de no pocos profesionales de la información.

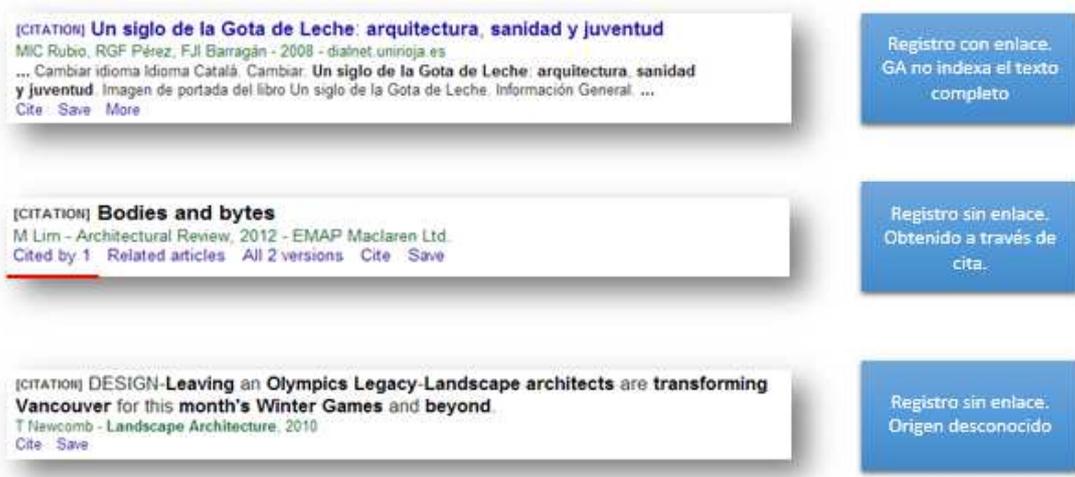


Ilustración 2: GA utiliza la palabra [Cita] para referirse a tipos de registros con orígenes muy diversos

Este problema encuentra su máxima expresión en aquellos registros que por un lado no ofrecen enlace externo y por otro tampoco parece que GA los haya recuperado a través de cita. Es en estos casos cuando parece que todas las puertas y ventanas de la transparencia se cierran definitivamente. Simplemente, se desconoce el origen de los metadatos. Para las tres muestras del estudio este tipo de registro representó el 28%, 50,42% y 32,61% de todos los trabajos recuperados. Para la muestra 3 además se pudo conocer su distribución por países. Es destacable que de los 48 registros de este tipo encontrados en la muestra, 18 pertenezcan a revistas de EEUU, 10 a revistas inglesas y 6 a alemanas. Es decir, entre los 3 países recuperan el 70,83% de este tipo de registros. La importancia para el caso de Alemania y EEUU es muy llamativa, alcanzando porcentajes del 66,66% y del 64,28% respectivamente.

En el Apéndice V se han analizado 15 de los 48 registros de este tipo que forman parte de la muestra 3. En concreto, se han seleccionado los 5 primeros para los países antes mencionados: EEUU, Alemania e Inglaterra. Se estudia la calidad de los registros (la referencia completa se obtiene a través del botón "cite") y se especula con las posibles fuentes de las que GA podría haberse ayudado.

En primer lugar se destaca la inexactitud y pobreza de sus metadatos. Así, para EEUU el primer registro presenta una sorprendente expresión, "THE CLIENT", antecediendo al título del artículo, el segundo carece de subtítulo, en el tercero no se indica el título

de la revista, mientras que para el cuarto falta el dato de paginación y el quinto aparece duplicado. Lo mismo sucede aunque en menor grado para el resto de la selección. Así, para el registro 3 y 5 de revistas inglesas no se indica el título de la revista, mientras que en el segundo registro de revistas alemanas precede al título la expresión "IM BLICKPUNKT" ("en el punto de mira") que podría ser el nombre de una sección fija de la revista.

También se trató de averiguar el origen de los metadatos. De los artículos estudiados, en 6 aparece alguna posible pista sobre la fuente utilizada en el registro de GA, aunque en un caso se trata aparentemente de una simple dirección incompleta (registro 3 de Inglaterra). Para el registro 3 de EEUU se buscó la expresión "*Science Publ Co Inc 117 West...*" en Google, refiriéndose probablemente a la editorial académica radicada en Singapur "World Scientific Publishing", aunque no se ha podido confirmar. Del mismo modo, para el registro 2 de Inglaterra tampoco se encontró evidencia suficiente. Para los registros 5 de EEUU y de Inglaterra las expresiones "*Hanley Wood*" y "*EMAP Business Publishing LTD*" hacen referencia a 2 empresas que se dedican a proveer, entre otras cosas, suscripciones a bases de datos y revistas relacionadas con la Arquitectura (en un caso también de otras áreas). Para el segundo caso, en concreto, la empresa publicita la revista "Architectural Review", a la que pertenece el artículo del registro. Para los dos casos se realizó una búsqueda a través del comando <site:>, válido para el buscador general de Google, por el título del trabajo sin resultados positivos. Por último, el registro 2 de EEUU presenta la expresión "*Natl Trust Historic...*" que hace referencia a la página web del "National Trust for Historic Preservation". Al realizar la siguiente búsqueda en Google <"counting down to done" site:http://www.preservationnation.org/> utilizando de nuevo el comando <site:> y el dominio de la página web efectivamente se encontró el artículo.

Nuevas preguntas surgen. ¿Existe una relación entre este tipo de registros y una baja cobertura de la disciplina? ¿Se insinúa un sesgo a favor de ciertos países (EEUU, Inglaterra, Alemania) al tratar de cubrir vacíos en la indexación de publicaciones conocidas aún a fuerza de saltarse los criterios de calidad que ellos mismos se imponen (exigencia de acceso al PDF o al menos al resumen de artículo)? ¿Por qué algunos registros parecen ofrecer pistas sobre el origen de los metadatos y otras no? Sin duda se requiere llevar a cabo una mayor investigación a través muestras mucho más extensas y específicas a este tipo de registros.

4. CONCLUSIONES.

En la primera parte del trabajo se han identificado serios problemas en la mayor parte de las metodologías utilizadas en los estudios de cobertura de GA:

- Se ha demostrado que el comando <site:> no permite evaluar el índice de GA al no tener en cuenta todos los enlaces externos (fuentes) del registro.
- Debido a la falta de control terminológico y al alto nivel de ruido documental que presenta GA, los estudios basados en búsquedas temáticas (“subject profile”) son de muy difícil aplicación por el número extraordinariamente alto de resultados que recupera y por la necesidad de analizar la pertinencia de cada trabajo recuperado.
- Los estudios llevados a cabo a nivel publicación (revista) se encuentran con los problemas metodológicos ya comentados por Pulgarín & Escalona (2007) en relación al diferente grado de indexación que una revista puede presentar en una base de datos, así como a problemas técnicos de muy difícil solución debidos principalmente a las limitaciones en la búsqueda avanzada de GA, a la falta de control de autoridades y al limitar la visualización de los registros a los primeros 1.000 resultados (Harzing, 2011).
- Se han utilizado el análisis de citas para el estudio del grado de similitud de GA respecto a WoS y Scopus, lo que permite calcular la tasa de solapamiento y los contenidos que recupera únicamente una base de datos en relación a las demás. Aunque algunos autores han utilizado estos análisis para valorar la extensión y profundidad del índice de GA, sin una limpieza y verificación de las citas, así como de una clasificación de éstas por tipología documental, es muy difícil alcanzar resultados relevantes (Jacsó, 2006; Rozear, 2008). Además, se ha encontrado que más del 12% de las obras citantes se pierden en GA al no indexar correctamente la bibliografía de estos trabajos (Meho & Yang, 2007).
- Los estudios basados en la contrastación con una lista de referencias (“bibliography method”) son la técnica más eficiente para el estudio de cobertura de GA, siempre que éste se realice contra una base de datos especializada y a través de muestras aleatorias con significación estadística. Los enfoques longitudinales, así como el análisis de los registros, de las fuentes utilizadas por GA o las lenguas y áreas geográficas cubiertas, enriquecen el análisis del índice y permiten conocer mejor su funcionamiento. Una versión

“enriquecida” de este modelo ha sido el utilizado en el estudio empírico llevado a cabo en la segunda parte del trabajo.

En relación al estado de la cuestión sobre la cobertura de GA, se ha identificado que:

- Una serie importante de áreas aún no han sido estudiadas suficientemente, principalmente en el área de humanidades, pero también para disciplinas concretas de ciencias sociales (ciencias políticas, derecho, etc.). GA presenta una alta cobertura de revistas científicas para el área de Ciencias, Tecnología y Ciencias de la Salud, siendo especialmente fuerte frente a WoS y Scopus en disciplinas como Matemáticas, Física, Ingenierías e Informática por el uso que los investigadores hacen para estas áreas de otras tipologías documentales (preprints, trabajos a congresos) y que GA indexa. En cambio, y a pesar de la mejora sustancial de su índice gracias a la indexación de editoriales como ACS a partir de los años 2008-2010, sigue ofreciendo una cobertura inferior para el área de Química.
- Para las Ciencias Sociales se ha encontrado que en general GA es capaz de recuperar un número muy superior de citas respecto a WoS o Scopus. En todo caso, estas bases de datos (WoS y Scopus) no deben ser consideradas recursos de referencia para esta área. Aún así, cuando el análisis se hace por tipologías documentales se encuentra que GA no recupera en cambio una parte sustancial de las revistas nucleares de estas disciplinas. Economía es una excepción en este caso, demostrando GA una buena cobertura de las revistas internacionales además de la recuperación de una gran cantidad de literatura gris (working papers, informes, etc.) gracias a la indexación de bases de datos en libre acceso como RePEC.
- Respecto a humanidades, los pocos estudios llevados a cabo ofrecen una baja tasa de cobertura y una indexación inconsistente de las principales publicaciones internacionales.

Respecto al estudio llevado a cabo para el área de Arquitectura, se encontraron tasas de cobertura muy bajas para los tres períodos (19,53%, 30,99% y 36,03% respectivamente). Además, la calidad del índice fue baja, siendo los registros sin enlace externo ni citas recibidas los más habituales para la muestra 2 y manteniendo un peso importante para las otras dos muestras.

El estudio confirma las ideas de White (2006). GA deja fuera de su índice a muchas editoriales científicas con uno o dos títulos, centrándose en un reducido número de editoriales científicas, las más importantes. Esto afecta a la capacidad de recuperación de literatura científica y profesional de ciertas disciplinas, sobre todo aquellas con una menor tradición científica y con un carácter poco internacional de sus publicaciones, como es el caso de Arquitectura. A pesar que éstas pueden solicitar su indexación al buscador científico de GA, no parece entrar esto dentro de los objetivos estratégicos de la mayor parte de las publicaciones, las cuales presentan en general un eminente carácter profesional.

En relación a la cobertura por áreas geográficas, se ha encontrado un fuerte desequilibrio entre los distintos países. Australia destaca con un 100% de cobertura gracias a la publicación de sus revistas en editoriales internacionales, a la existencia de una base de datos nacional como Informit y a una actividad de difusión activa por parte de los autores a través del uso de los repositorios institucionales y las redes sociales científicas. Desde el punto de vista negativo destaca Alemania con sólo un 22% de registros recuperados. De ellos, el 66% fueron a través de los registros de peor calidad posible (sin enlace externo ni citas recibidas). En general, y debido a que la mayor parte de los registros se han recuperado desde bases de datos de acceso libre, se confirma la hipótesis de Kousha, K., & Thelwall (2007) de que GA presenta un cierto sesgo favorable hacia aquellos países que tienen bases de datos nacionales para la difusión científica (como CNKI para China o Dialnet para España).

Finalmente, el brusco cambio que ha sufrido el sistema de evaluación académica a nivel global, obliga al investigador a buscar ahora la mayor visibilidad posible de su producción científica. De esta forma, se abren nuevos caminos ya vislumbrados en la muestra 3 que permiten pensar en un incremento sustancial de la indexación por parte de GA de la literatura de Arquitectura, principalmente gracias a un mayor uso de los repositorios institucionales, de las redes sociales científicas tipo ResearchGate o Academia.edu, así como en un cambio en la estrategia editorial en busca de un mayor rigor científico por parte de las revistas del área.

BIBLIOGRAFÍA

- Adamick, J., & Reznik-Zellen, R. (2010). Representation and Recognition of Subject Repositories. *D-Lib Magazine*, 16(9/10). doi:10.1045/september2010-adamick
- Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, 30(1), 60–81. Recuperado de http://scholarworks.montana.edu/xmlui/bitstream/handle/1/3193/Arlitsch-Obrien-LHT-GS-final-revised_2012-02-18.pdf?sequence=1
- Bar-Ilan, J. (2010). Citations to the «Introduction to informetrics» indexed by WOS, Scopus and Google Scholar. *Scientometrics*, 82(3), 495–506.
- Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century – A review. *Journal of Informetrics*, 2(1), 1–52.
- Lasda Bergman, E. M. (2012). Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar. *Journal of Academic Librarianship*, 38(6), 370-379. doi:10.1016/j.acalib.2012.08.002
- Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Medical Research Methodology*, 13(1), 131. doi:10.1186/1471-2288-13-131
- Bohannon, J. (2014). Google Scholar Wins Raves—But Can It Be Trusted? *Science*, 343(6166), 14–14.
- Bosman, J., van Mourik, I., Rasch, M., Sieverts, E., & Verhoeff, H. (2006). Scopus reviewed and compared: The coverage and functionality of the citation database Scopus, including comparisons with Web of Science and Google Scholar. Recuperado a partir de <http://igitur-archive.library.uu.nl/uul/2006-1220-200432/UUindex.html>
- Bramer, W. M., Giustini, D., Kramer, B. M., & Anderson, P. F. (2013). The comparative recall of Google Scholar versus PubMed in identical searches for biomedical systematic

- reviews: a review of searches used in. *Systematic Reviews*, 2(1), 1-9. doi:10.1186/2046-4053-2-115
- Brantley, P. (3 de julio de 2007). Science Direct-ly into Google. *TOC O'Reilly*. Recuperado a partir de <http://toc.oreilly.com/2007/07/science-directly-into-google.html>
- Brewerton, A. (2012). New ways of supporting researchers. *CILIP Update*, (February).
Recuperado a partir de
http://wrap.warwick.ac.uk/42031/1/WRAP_Brewerton_Update_02_pp40-42.pdf
- Burright, M. (2006). Google Scholar: Science & Technology. *Issues in Science and Technology Librarianship*, (Winter). Recuperado a partir de <http://www.istl.org/06-winter/databases2.html>
- Chen, X. (2010a). The declining value of subscription-based abstracting and indexing services in the new knowledge dissemination era. *Serials Review*, 36(2), 79-85.
- Chen, X. (2010b). Google Scholar's Dramatic Coverage Improvement Five Years after Debut. *Serials Review*, 36(4), 221-226. doi:10.1016/j.serrev.2010.08.002
- Christianson, M. (2007). Ecology articles in Google Scholar: levels of access to articles in core journals. *Issues in Science and Technology Librarianship*, 49. Recuperado a partir de http://www.istl.org/07-winter/refereed.html?a_aid=3598aabf
- Cusker, J. (2013). Elsevier Compendex and Google Scholar: A Quantitative Comparison of Two Resources for Engineering Research and an Update to Prior Comparisons. *The Journal of Academic Librarianship*, 39(3), 241-243. doi:10.1016/j.acalib.2013.02.001
- De Groote, S. L., & Raszewski, R. (2012). Coverage of Google Scholar, Scopus, and Web of Science: A case study of the h-index in nursing. *Nursing Outlook*, 60(6), 391-400. doi:10.1016/j.outlook.2012.04.007
- Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google scholar experiment: How to index false papers and manipulate bibliometric indicators.

Journal of the Association for Information Science & Technology, 65(3), 446-454.

doi:10.1002/asi.23056

Dobbs, D. (12 de Abril de 2013). When the rebel alliance sells out. *The New Yorker*.

Recuperado a partir de

<http://www.newyorker.com/online/blogs/elements/2013/04/elsevier-mendeley-journals-science-software.html>

Forsyth, A. (2007). The rise of the nerds? Interdisciplinary research and architecture.

Archnet-IJAR, International Journal of Architectural Research, 1(3), 177-182.

Gardner, S., & Eng, S. (2005). Gaga over Google? Scholar in the social sciences. *Library Hi*

Tech News, 22(8), 42-45.

Gehanno, J.-F., Rollin, L., & Darmoni, S. (2013). Is the coverage of Google Scholar enough to

be used alone for systematic reviews. *BMC medical informatics and decision making*, 13(1), 7.

Giustini, D., & Barsky, E. (2005). A look at Google Scholar, PubMed, and Scirus:

comparisons and recommendations. *Journal of the Canadian Health Libraries Association*, 26(3), 85-89.

Giustini, D., & Boulos, M. N. K. (2013). Google Scholar is not enough to be used alone for

systematic reviews. *Online journal of public health informatics*, 5(2), 214.

Gray, J. E., Hamilton, M. C., Hauser, A., Janz, M. M., Peters, J. P., & Taggart, F. (2012,

Summer). Scholarish: Google Scholar and its Value to the Sciences. *Issues in Science & Technology Librarianship*, (70), 1-3.

Harzing, A.-W., & Van Der Wal, R. (2009). 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*, 60(1), 41-46.

Harzing, A.-W. (2011). *The Publish or Perish Book: Your guide to effective and responsible citation analysis*. Melbourne (Australia): Tarma Software Research.

- Harzing, A.-W. (2013). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics*, 94(3), 1057-1075.
doi:10.1007/s11192-012-0777-7
- Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is Google Scholar? A comparison to library databases. *College & Research Libraries*, 70(3), 227-234.
- Inger, S., & Gardner, T. (2012). *How Readers Discover Content in Scholarly Journals*. Renew Training.
- Jacsó, P. (2005a). 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-BANGALORE-*, 89(9), 1537. Recuperado de <http://choo.fis.utoronto.ca/FIS/courses/LIS1325/Readings/jacso.pdf>
- Jacsó, P. (2005b). Google Scholar: the pros and the cons. *Online information review*, 29(2), 208-214. Recuperado de http://www.researchgate.net/publication/220207633_Google_Scholar_the_pros_and_the_cons/file/3deec528451d9a5d5a.pdf
- Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online information review*, 30(3), 297-309.
- Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, 34(1), 175-191.
- Jones, D. Y. (2005). Biology article retrieval from various databases: making good choices with limited resources. *Issues in Science and Technology Librarianship*, 44. Recuperado a partir de <http://www.istl.org/05-fall/refereed.html>
- Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58(7), 1055-1065.

- Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics*, 74(2), 273-294.
- Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62(11), 2147-2164.
- Kreisman, R. (2013). Thomson Reuters-Google Scholar Linkage Offers Big Win for STM Users and Publishers. *Outsell*. Recuperado a partir de http://wokinfo.com/media/pdf/outsell_trgs_reciprocallinking
- Lewandowski, D. (2010). Google Scholar as a tool for discovering journal articles in library and information science. *Online Information Review*, 34(2), 250-262.
- Levine-Clark, M., & Kraus, J. (2007). Finding chemistry information using Google Scholar: a comparison with Chemical Abstracts Service. *Science & technology libraries*, 27(4), 3-17.
- Mayr, P., & Walter, A.-K. (2008). Studying journal coverage in Google Scholar. *Journal of Library Administration*, 47(1-2), 81-99.
- Meho, L. I., & Yang, K. (2007). 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*, 58(13), 2105-2125.
- Meier, J. J., & Conkling, T. W. (2008). Google Scholar's coverage of the engineering literature: An empirical study. *The Journal of Academic Librarianship*, 34(3), 196-201.
- Mingers, J., & Lipitakis, E. A. (2010). Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics*, 85(2), 613-625.
- Neuhaus, C., Neuhaus, E., Asher, A., & Wrede, C. (2006). The depth and breadth of Google Scholar: An empirical study. *portal: Libraries and the Academy*, 6(2), 127-141.

- Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics*, 1(2), 161-169.
doi:10.1016/j.joi.2006.12.001
- Notess, G. R. (2005). Scholarly web searching: Google Scholar and Scirus. *Online*, (Jul/ Aug).
Recuperado a partir de <http://www.infotoday.com/online/jul05/OnTheNet.shtml>
- Orduna-Malea, E., Ayllon, J. M., Martin-Martin, A., & Lopez-Cozar, E. D. (2014). Empirical Evidences in Citation-Based Search Engines: Is Microsoft Academic Search dead? *arXiv preprint*. Recuperado a partir de <http://arxiv.org/abs/1404.7045>
- Pulgarín Guerrero, A. (2007). Medidas del solapamiento en tres bases de datos con información sobre ingeniería. *Anales de documentación*, (10), 335-344.
- Ramos Carranza, A., & Añón Abajas, R. M. (2012). La difusión como proyecto de investigación: revistas de investigación en proyectos arquitectónicos. En *4IAU 4ª Jornadas Internacionales sobre Investigación en Arquitectura y Urbanismo, Valencia, 2011*.
- Rozear, H. M. (2009). Where Google Scholar stands on art: an evaluation of content coverage in online databases. *Art Libraries Journal*, 34(2), 21-25.
- Salisbury, L., & Tekawade, A. (2006). Where Is Agricultural Economics and AgriBusiness Research Information Published and Indexed? A Comparison of Coverage in Web of Knowledge, CAB Abstracts, EconLit, and Google Scholar. *Journal of Agricultural & Food Information*, 7(2/3), 125-143. doi:10.1300/108v07n0210
- Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association: JMLA*, 95(4), 442.
- Sorli, Á., & Mochón, G. (2012). Revistas españolas de arquitectura, ciencias de la construcción y urbanismo: visibilidad e internacionalidad. Recuperado a partir de <http://ignucius.bd.ub.edu:8180/jspui/handle/123456789/776>

- Sumner, N. (July 5-7). Developing the digital researcher. En *ICICTE 2012 Proceedings*.
Rhodes, Greece. Recuperado a partir de
<http://www.icicte.org/Proceedings2012/Papers/05-1-Sumner.pdf>
- Technopolis Group. (2011). *Review Architecture and the Built Environment TU/e and TU Delft*.
Recuperado a partir de
https://www.tue.nl/fileadmin/content/onderzoek/Kwaliteitstoets_onderzoek/OZ_2003-2009_TUE_Built_Environment.pdf
- Torres-Salinas, D., & Delgado-López-Cózar, E. (2009). Estrategia para mejorar la difusión de los resultados de investigación con la Web 2.0. *El profesional de la información*, 18(5), 534-539.
- Viñas, A., & Minobis, E. (2007). La evaluación de las revistas en la investigación en arquitectura y urbanismo. En *XVIII Jornadas ABB: las publicaciones periódicas de arquitectura, construcción y urbanismo*. León. Recuperado a partir de
<http://www.bibarquitectura.us.es/nuevapag/evaluacion/comunicacion.pdf>
- Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. *Information processing & management*, 43(4), 1121-1132.
- Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *portal: Libraries and the Academy*, 9(1), 5-24.
- Walters, W. H. (2011). Comparative Recall and Precision of Simple and Expert Searches in Google Scholar and Eight Other Databases. *portal: Libraries and the Academy*, 11(4), 971-1006. doi:10.1353/pla.2011.0042

ANEXOS

ANEXO I

Cronología de Google Académico

17 de noviembre de 2004: “Google Scholar”. El buscador académico es lanzado oficialmente. Curiosamente, su lanzamiento se produce sólo pocos días después de la puesta en el mercado de otra base de datos multidisciplinar, Scopus, de Elsevier.

Mayo 2005: “Library links” (en un principio llamado “**Google Scholar Institutional Access**”). Permite que los usuarios de las bibliotecas, tanto dentro del campus como desde fuera, puedan conocer y tener acceso a los recursos electrónicos suscritos por su institución que aparezcan dentro de los resultados de búsqueda.

http://googlepress.blogspot.com.es/2005/05/google-scholar-institutional-access_10.html

<http://googleblog.blogspot.com.es/2005/05/library-access.html>

Febrero 2006: “Local libraries”. Desde la creación de GS se había cerrado un acuerdo con la OCLC para y a través de su catálogo colectivo WorldCat ofrecer la posibilidad de localizar la biblioteca más cercana al usuario que ofreciese el documento consultado (gracias a un link dispuesto en la parte inferior de cada registro). Es a partir de esta fecha, febrero 2006, cuando esta posibilidad se abre para cualquier biblioteca del mundo. Respecto a nuestro país, GS tiene firmado un acuerdo con Rebiun para este fin, lo que permite con un solo click reenviar la búsqueda al propio catálogo colectivo de ésta. Para acceder al enlace hay que clicar, debajo de cada registro, en “más” (more). El enlace a Rebiun sólo aparece en caso de que el documento se halle en el propio catálogo.

<http://googleblog.blogspot.com.es/2006/02/global-searches-go-to-local-libraries.html>

20 abril de 2006: “Recent articles”. Con esta nueva posibilidad de búsqueda GS trata de responder a las continuas críticas debido a la imposibilidad de ordenar los

resultados por fecha, viéndose empujados hacia abajo artículos recientes pertinentes pero que aún no han tenido la posibilidad de ser citados. De todas formas, como se indica en el propio blog de GS, no se trata realmente de un filtro por fecha a secas, sino que también se tiene en cuenta la importancia del autor, cuántas citas tiene, cuándo fue escrito, etc.

<http://googleblog.blogspot.com.es/2006/04/keeping-up-with-recent-research.html>

Agosto 2006: “Related Articles”. Link disponible debajo de cada resultado. Da acceso a los artículos indexados por GS que tienen una mayor relación temática con el documento seleccionado. La ordenación de estos se realiza por el grado de similitud con el trabajo original, pero también por la relevancia de cada publicación. Se dispone a partir de esta fecha de una nueva forma de navegar por los resultados, además de por “cited by”.

<http://googleblog.blogspot.com.es/2006/08/exploring-scholarly-neighborhood.html>

Diciembre 2006: “Google Patent Search”. El 13 de diciembre de 2006 Google lanzaba su propio buscador de patentes. Sólo unos días después, los usuarios del buscador académico se sorprendían al encontrar el contenido de este proyecto integrado en los resultados de búsqueda de GS.

<http://googleblog.blogspot.com.es/2006/12/now-you-can-search-for-us-patents.html>

<http://www.seobythesea.com/2006/12/google-patent-search-results-added-to-google-scholar/>

Verano 2007: “Elsevier”. Desde el lanzamiento de GS, Elsevier, quizá la mayor editorial académica del mundo, no había permitido que la araña de Google indexase sus contenidos por considerarla un competidor directo a sus productos. De hecho, era la única gran editorial (con la excepción de ACS, de un tamaño muy inferior) que le había dado la espalda al gigante Google. Finalmente, y a pesar de tener productos rivales directos como Scirus o Scopus, la gran editorial holandesa llega a un acuerdo con Google y permite que las revistas de su base de datos Science Direct (controla directamente su copyright) sean indexadas por la compañía de California. Aunque GS ya indexaba parte de las revistas de Elsevier indirectamente (a través de agregadores como Ingenta) sin duda es un gran triunfo para Google.

<http://toc.oreilly.com/2007/07/science-directly-into-google.html>

Noviembre de 2009: “Case law”, que permite realizar búsquedas sobre la jurisprudencia existente para un tema concreto en EEUU. En concreto, los dictámenes de la Corte Suprema de EEUU desde 1791, así como de los juzgados federales, apelación, impuestos y bancarrota desde 1923; y apelaciones del estado y de la corte suprema desde 1950. Por desgracia esta funcionalidad no permite la consulta homóloga a los casos legales de otros países.

<http://googleblog.blogspot.com.es/2009/11/finding-laws-that-govern-us.html>

https://www.sfbar.org/articles/google_scholar.aspx

<http://tarltonguides.law.utexas.edu/content.php?pid=242166&sid=1999301>

<http://jamesonlawlibrary.wordpress.com/2012/09/23/new-features-on-google-scholar-to-help-you-locate-case-law/>

Junio 2010: “Google Scholar Alerts”. Permite crear alertas, bien para una búsqueda concreta, bien respecto a nuevas citas a un artículo concreto.

<http://googlescholar.blogspot.com.es/2010/06/google-scholar-alerts.html>

Julio 2011: “Google Scholar Citations”.

<http://googlescholar.blogspot.com.es/2011/07/google-scholar-citations.html>

Abril 2012: “Google Scholar Metrics”.

<http://googlescholar.blogspot.com.es/2012/04/google-scholar-metrics-for-publications.html>

Junio 2012: “Sort by date”. Aunque ya a partir de 2006 GS habilitó alguna funcionalidad para tratar de recuperar los artículos citados más recientemente, no es

hasta esta fecha cuando GS permite finalmente escoger el orden de los resultados, bien por relevancia o bien por fecha. De todas formas, un año después aún se pueden leer quejas en el foro de Google sobre las disfuncionalidades de esta nueva característica. Un usuario denuncia que al clicar sobre esta funcionalidad, GS lo que realmente hace es eliminar todos los registros de años anteriores al presente.

<http://www.lawtechnologytoday.org/2012/07/big-changes-at-google-scholar/>

<http://googlescholar.blogspot.com.es/2012/06/search-recent-additions-sorted-by-date.html>

<https://productforums.google.com/forum/#!msg/websearch/P1z4qMRYylg/jzI9019oKWcJ>

Noviembre 2013: “Google Scholar Library”

<http://googlescholar.blogspot.com.es/2013/11/google-scholar-library.html>

Enero 2014: “Scirus anuncia el cierre”.

Febrero 2014: “Acuerdo Google Scholar vs Thomson Reuters”

Abril 2014: “Is Microsoft Academic Search dead”. El 29/04/2014 el grupo de investigación granadino EC3 pone en marcha un nuevo blog dedicado a GS. Su primer post, único publicado hasta el momento, presenta un estudio empírico en el que muestra una terrible falta de actualización del motor de búsqueda académico: frente a los más de 2 millones de documentos subidos a su base de datos en 2010, sólo 8,147 fueron cargados en 2013 y tan sólo 802 en 2014. Si se confirma la muerte de Microsoft Academic Search, GS quedaría en el mercado como el único buscador científico generalista gratuito, sin ningún tipo de competencia.

<http://googlescholar Digest.blogspot.com.es/2014/04/empirical-evidences-microsoft-academic-search-dead.html>

ANEXO II

MUESTRA 1

Años cubiertos: 1950-1989

Nº de registros en Avery 1990-2009: 209.804

Rango números de acceso: 63.987-295.120

Tamaño de la población: 231.133 registros

Tamaño de la muestra: 384 (nivel de confianza e intervalo: 95%, 5)

Determinación aleatoria de la muestra: <http://www.random.org/integers/>

MUESTRA 2

Años cubiertos: 1990-2009

Nº de registros en Avery 1990-2009: 292.789

Rango números de acceso: 295.120 – 654.500

Tamaño de la población: 359.380 registros

Tamaño de la muestra: 384 (nivel de confianza e intervalo: 95%, 5)

Determinación aleatoria de la muestra: <http://www.random.org/integers/>

MUESTRA 3

Años cubiertos: 2010 – 2014 (hasta junio)

Nº de registros 2010-2014: 53.834 registros (a 2 de julio de 2014)

Rango números de acceso: 643.500 – 712.114

Tamaño de la población: 68.614 registros

Tamaño de la muestra: 382 (nivel de confianza e intervalo: 95%, 5)

Determinación aleatoria de la muestra: <http://www.random.org/integers/>

ANEXO III

“The depth and breadth of Google Scholar: An empirical study” (Neuhaus et al., 2006)
<p><u>1er criterio:</u> se evalúa el contenido de GA respecto a 47 bases de datos que los autores clasifican en 7 áreas del conocimiento (Economía y negocios, Educación, Humanidades, Ciencia y medicina, Ciencias sociales y Multidisciplinar). Entre ellas podemos encontrar prestigiosas bases de datos como ABI, ERIC, Historical Abstracts, ACM Digital Library, Biological Abstracts, GeoRef, PubMed, PsycInfo, etc., además de los portales de algunos de los principales editores científicos.</p>
<p><u>2º criterio:</u> para cada uno de las bases de datos se seleccionó una muestra aleatoria de 50 registros.</p>
<p><u>Valoración:</u> trabajo de referencia en este tipo de estudios, es tomado por Chen como ejemplo metodológico a seguir. De todas formas, y a pesar de que se comprueban más de 4.000 registros, el tamaño de la muestra para cada base de datos es muy pequeño (50 registros), lo no parece permitir una validación estadística. Existen aplicaciones online que permiten calcular el tamaño óptimo de una muestra (véase por ejemplo, http://es.gmi-mr.com/solutions/sample-size-calculator.php). Así por ejemplo, para un nivel del 95% de confianza con un intervalo de confianza del 5% (mínimo exigible), para poblaciones grandes se exige una muestra de 384 ítems. Este estudio no cumple pues el segundo requisito.</p>

“Google Scholar coverage of a multidisciplinary field” (Walters W.H., 2007)
<p><u>1er criterio:</u> se intenta identificar los mejores artículos publicados entre 1990 y 2000 para la línea de investigación “later-life migration” mediante la búsqueda en distintas bases de datos, consulta manual de revistas y revisión de las referencias bibliográficas los</p>
<p><u>2º criterio:</u> el tamaño de la selección es de 155 artículos</p>
<p><u>Valoración:</u> la construcción de la lista de referencia es poco transparente y el número total de ítems bajo. Parece que este estudio valore más la capacidad de GA de indexar trabajos de alta calidad en la línea de investigación que la de testar la capacidad de recoger en su base de datos las publicaciones del área. El sesgo se debe a que, gracias a que GA recupera trabajos de forma indirecta a través de las citas en</p>

terceros trabajos, siempre tendrá más posibilidades de recuperar trabajos con un fuerte impacto, de ahí que el estudio no sea del todo representativo del conjunto del área de estudio. Esto se podría haber paliado parcialmente si se hubiese analizado el origen de cada registro dentro del índice de GA (a través del editor de la revista, bases de datos, repositorios, o bien citas).

“Comparing alternatives to the Web of Science for coverage of the social sciences’ literature” (Norris M. & Oppenheim C., 2007)

1er criterio: los 33.533 artículos presentados al sistema de evaluación nacional en UK (RAE) para el año 2001 para el área de Ciencias Sociales. También se incluyó en el estudio un conjunto de 581 revistas en idioma no inglés de la lista “International Bibliography of the Social Science” (IBSS) con el objeto de identificar sesgos en la cobertura por el idioma.

2º criterio: de los dos conjuntos anteriores se tomó una única muestra aleatoria de 306 artículos.

Valoración: estudio muy completo ya que realiza una evaluación de la cobertura a nivel revista, a nivel registro y a nivel citas. Para el análisis a nivel registro diferencia para GA si son tipo cita o si por el contrario han sido recuperados directamente por el buscador. Cumple con los dos criterios mencionados, aunque la muestra podría haber sido algo más grande. Se comparan los resultados alcanzados con las bases de datos WoS, Scopus y CSA Illumina.

“Finding chemistry information using Google Scholar: a comparison with Chemical Abstracts Service” (Levine-Clark M. & Kraus J., 2007)

1er criterio: Base de datos Chemical Abstracts Service (CAS)

2º criterio: la lista de registros se obtienen a partir de 5 búsquedas realizadas en CAS: 2 temáticas, 2 por componentes y una por autor (se lanzaron las mismas búsquedas en GA). En CAS se recuperaron 316 resultados.

Valoración: el estudio cumple con el primer criterio, ya que CAS una de las bases de datos de referencia para el área de Química. Pero en cambio, no cumple con el segundo criterio, ya que la selección de la muestra no es aleatoria, lo que desvirtúa en cierto grado la representatividad sobre el conjunto de los registros de la base de datos. El tamaño de la muestra es algo inferior al deseado para un estudio

estadístico con un nivel de confianza del 95% y con un intervalo de confianza del 5%.

“Ecology articles in Google Scholar: levels of Access to articles in core journals”
(Christianson M., 2007)

1er criterio: se tienen en cuenta para el estudio las revistas comunes a dos listas diferentes: “Ecology Abstracts” de CSA (Cambridge Scientific Abstracts) y las revistas de ecología incluidas en el JCR de WoS. Los títulos se buscaron en WoS recuperando un total de 86.774 trabajos entre los años 1945-2005, conformando la población de estudio.

2º criterio: muestra aleatoria de 840 registros.

Valoración: el estudio cumple con los dos criterios de Chen. Además, los registros recuperados se clasificaron mediante la siguiente tipología: acceso directo al texto completo, con acceso completo previo pago, con abstract, con cita completa, cita incompleta pero con link a la referencia completa, y cita incompleta. También se identificaron los proveedores de información principales de GA.

“Where Google Scholar stands on art: an evaluation of content coverage in online databases” (Rozeat, 2009)

1er criterio: La obra de referencia “Dictionary of Art Historians”.

2º criterio: De todos los historiadores que aparecen en la obra se selecciona a 12 autores en base a una serie de criterios (idioma anglosajón, nacimiento posterior a 1925, etc.). La muestra final está formada por las obras de estos 12 autores y se limita a aquellas tipologías documentales típicas de una revista (artículos, editoriales, revisiones de libros, obituarios, etc.) Un total de 472 publicaciones.

Valoración: la selección de la muestra es original aunque hasta cierto punto queda la duda si estas obras son finalmente representativas del área de estudio. El tamaño de la muestra es suficiente. Lo interesante de este trabajo es que los resultados de cobertura de GA se comparan con dos bases de datos especializadas y otra más generalista: The Bibliography of the History of Art (BHA) y Art Full Text/art Index Retrospective, así como A&H de WoS.

“Google Scholar’s coverage of the engineering literature: An empirical study” (Meier

J.J. & Conkling T.W., 2008)
<u>1er criterio:</u> base de datos Compendex
<u>2º criterio:</u> se realizan 8 búsquedas representativas de distintas áreas de la ingeniería para cada década desde 1950 a 2000. De los resultados de cada búsqueda se toman muestras aleatorias de 20 registros. En total, se verificaron 960 trabajos.
<u>Valoración:</u> se trata de un estudio longitudinal que cubre más de 6 décadas. El enfoque permite ir descubriendo los diferentes grados de cobertura de GA según el año de publicación. El único interrogante es su validez estadística, ya que aunque el número de registros total es elevado, las muestras aleatorias de 20 registros pueden no ser suficientes si la pregunta ha devuelto un número alto de resultados.

“Google Scholar as a tool for discovering journal articles in library and information science” (Lewandowski D., 2010)
<u>1er criterio:</u> analiza 35 revistas de lengua inglesa de máxima relevancia internacional (todas ellas JCR).
<u>2º criterio:</u> se analizan todos los artículos publicados por aquellas entre los años 2004-2006. En total, 3.580 artículos.
<u>Valoración:</u> Este estudio, a pesar del tamaño de la muestra, sólo puede ser relevante para conocer el grado de cobertura de GA sobre las revistas de máxima relevancia y carácter internacional del área de estudio, pero no respecto a una revista de nivel medio. Por ello, este estudio sólo puede tener un carácter complementario a aquellos trabajos que realizan estudios frente a una base de datos de referencia. Por otra parte, se limita el análisis a un período muy concreto y favorable a GA (no son trabajos antiguos, pero tampoco lo suficientemente actuales para que no puedan ser recuperados vía cita). Se realiza un estudio de los registro según el grado de acceso y a la tipología del editor.

“Google Scholar’s Dramatic Coverage Improvement Five Years after Debut” (Chen X., 2010b)
<u>1er criterio:</u> 7 bases de datos ya estudiadas en el trabajo de Neuhaus et al. (2006) y el portal de la editorial ACS.
<u>2º criterio:</u> muestras aleatorias de 50 registros cada una.
<u>Valoración:</u> se trata de una réplica del trabajo de Neuhaus et al. (2006). El objeto de

estudio es determinar si el grado de indexación de bases de datos que permiten el acceso libre a las arañas de Google se ha incrementado en los últimos 5 años. Este tipo de trabajos son idóneos para conocer la evolución del índice de GA. Las fortalezas y debilidades metodológicas son las mismas al trabajo original.

“Elsevier Compendex and Google Scholar: A quantitative comparison of two resources for Engineering research and an update to prior comparisons” (Cusker J., 2013)

1er criterio: base de datos Compendex

2º criterio: se utiliza la misma metodología del trabajo anterior, pero en vez de tomar una muestra aleatoria de 20 registros de cada búsqueda, la investigación se realiza sobre los 20 primeros registros.

Valoración: el interés de un trabajo de cobertura que replica la metodología de uno anterior es que permite conocer el crecimiento retrospectivo del índice. Respecto a los criterios de Chen, este estudio comparte las mismas dudas del anterior a lo que ahora hay que sumar que las muestras no son aleatorias.

“Is the coverage of Google Scholar enough to be used alone for systematic reviews” (Gehanno JF, Rollin L & Darmoni S, 2013).

1er criterio: Trabajos que forman parte de revisiones sistemáticas en medicina.

2º criterio: se buscan en Medline revisiones sistemáticas publicadas entre 2008-2009, filtradas por JAMA y Cochrane en las que se hayan consultado al menos dos bases de datos y la búsqueda no estuviese limitada al idioma inglés. Se encuentran 29 publicaciones. Se toma como muestra de estudio las referencias bibliográficas incluidas en estos trabajos. Un total de 733 publicaciones.

Valoración: El trabajo presenta un problema metodológico inherente al propósito del estudio. El objeto es conocer si GA puede ser una fuente idónea para realizar revisiones sistemáticas, es decir, en qué grado indexa trabajos que potencialmente pueden ser usados en este tipo de publicaciones. El problema reside en que al seleccionar la muestra a partir de las referencias bibliográficas de trabajos ya publicados, la posibilidad de que GA indexe éstos, sobre todo de forma indirecta mediante el registro tipo “cita”, se incrementan exponencialmente. Además, el trabajo no se acompaña de una cuantificación del tipo de la procedencia de los

registros de GA, por lo que se desconoce hasta qué punto este factor ha podido alterar los resultados del estudio.

“Google Scholar is not enough to be used alone for systematic reviews” (Giustini D. & Bouos M.N.K., 2013)

1er criterio: Trabajos que forman parte de revisiones sistemáticas en medicina.

2º criterio: sólo se tiene en cuenta una revisión sistemática. Se toma como muestra el total de trabajos incluidos en su bibliografía. En total, 476 artículos.

Valoración: se trata de una réplica directa del trabajo anterior. A los problemas metodológicos ya comentados se suma el criterio arbitrario en la selección de la única revisión sistemática tenida en cuenta.

ANEXO IV

País	nº art	En GA %	PDF free	PDF inst.	Ref.	Sin link	Cita	Editor	BDs	Rep. Inst.	Pág. Univ.	Pág. Autor	Red Social
Inglaterra	89	37 42%	8	9	7	10	3	18	14	7	2	2	2
EEUU	92	28 30%	2	3	5	18		4	6	2			
Alemania	41	9 22%			3	6		1	3				
España	27	15 56%	5		10			5	14	4			
Italia	19	8 42%			6	2		2	6				
Canadá	13	2 15%			1	1		1					
China	10	7 70%			7				7				
Japón	10	1 10%				1							
Corea	9					2							
Australia	8	8 100%	3	1	4			7	3	3	1		1

Holanda	7	1	14%				1												
México	6	2	33%	1		1			1	1									
República Dominicana	5	2	40%			2				2									
Chile	5	4	80%	2		2			1	3									
Venezuela	2	0	0%																
India	4	2	50%				2												
Argentina	3	2	67%			2				2									
Perú	4	0	0%																
Portugal	4	3	75%			3				3									
Colombia	4	1	25%			1			1	1									
Dinamarca	3	1	33%				1												
Sudáfrica	3	0	0%																
Irlanda	2	0	0%																
Francia	2	1	50%				1												
Finlandia	2	1	50%				1												
Eslovenia	2	2	100%	2					1	2									
Noruega	1	0	0%																
Suecia	1	0	0%																
Israel	1	0	0%																
Nueva Zelanda	1	0	0%																
Croacia	1	0	0%																

ANEXO V

Selección de registros sin enlace ni citas recibidas (muestra 3)

EEUU

[CITATION] **THE CLIENT The Caution on Campus** Higher-ed institutions are having a harder time pushing projects through budget barriers, but that doesn't mean ...
E Beck - Landscape Architecture, 2011
[Cite](#) [Save](#)

No existen indicios del origen de los metadatos. Previo al título originario en Avery aparece en el registro de GA y en mayúsculas las palabras "THE CLIENT". Parece un error.

[CITATION] **Counting Down to Done**
LM Roberts - PRESERVATION, 2010 - NATL TRUST HISTORIC ...
[Cite](#) [Save](#)

La web de "Natl Trust Historic..." ¿posible origen de los metadatos? No aparece el subtítulo que sí se encontraba en el registro de Avery. Al realizar la búsqueda por la expresión en Google lleva a la página web del "National Trust for Historic Preservation". Al realizar la siguiente búsqueda en Google <"counting down to done" site:http://www.preservationnation.org/> utilizando el comando <site:> y el dominio de la página web de la institución efectivamente se ha encontrado el artículo anterior.

[CITATION] **The Encyclopedia of Housing**
EJ Mueller - 2013 - ... SCIENCE PUBL CO INC 117 WEST ...
[Cite](#) [Save](#)

Metadatos pobres. No aparece el título de la revista en la referencia ("Journal of Architectural & Planning Research"). ¿"Science Publ Co Inc 117 West..." puede ser una pista sobre la fuente del registro? Se ha buscado la expresión en Google y es probable que se pueda referir a la editorial académica radicada en Singapur "World Scientific Publishing", aunque no se ha podido confirmar.

[CITATION] **You Can Make Your House Feel More Simple**-Interior design by Jean Larette
L Cregan - House Beautiful, 2010
Cite Save

En la referencia no se indica información sobre la paginación. No existe indicación del posible origen de los metadatos.

[CITATION] **Death by Cliché!** In Chip Kidd's latest graphic novel, Batman: Death by Design, the Caped Crusader and the Joker are joined by a cast of architectural ...
K Capps - Architect-Northbrook, 2012
Cite Save More

[CITATION] **DEATH BY CLICHE!**
K Capps - ARCHITECT, 2012 - HANLEY WOOD, LLC ONE ...
Cite Save

Registro duplicado. En el primer caso no aparece indicación de origen. El título de la revista parece contener un error. En el segundo caso, no aparece el subtítulo en el registro.

¿"Hanley Wood" puede ser una pista del origen de los metadatos? A través de la correspondiente búsqueda en Google se llega a la página web de la empresa "Hanley Wood" la cual ofrece servicios de "información, prensa, creación de eventos y marketing para empresas dedicadas a la arquitectura y construcción tanto residencial como comercial", <http://www.hanleywood.com/>. La búsqueda en cambio < death by cliché site:http://www.hanleywood.com/> no ha obtenido resultado alguno.

ALEMANIA

[CITATION] **Wachstum, Flut und Fluch**-Dhaka und Lagos: Dichte in gefährdeten Regionen
M Zettel - Archithese, 2011
Cite Save

Se desconoce el origen del registro. Se añade una segunda parte del título que no se recoge en el registro de Avery.

[CITATION] **IM BLICKPUNKT EXPO 2010: ZWISCHEN BANALITÄTEN UND MAGIE**
H Adam - DB: Deutsche Bauzeitung, 2010
[Cite](#) [Save](#)

Se desconoce el origen de los metadatos. Se antecede en GA al título en Avery con las palabras “IM BLICKPUNKT” que más parece el nombre de una posible sección fija de la revista.

[CITATION] **Augmented Architecture** Über den **sinnlichen Umgang** mit **digitalen Medien** in der **Architektur**
U Hirschberg - Werk Bauen und Wohnen, 2011
[Cite](#) [Save](#)

GA no ofrece pistas de a través de qué dominio se rastrearon los metadatos del registro. La referencia es correcta.

[CITATION] **Wohnungsbau der grossen Zahl**-Ein Spaziergang durch die Genfer Ensembles der 1950er, 60er und 70er Jahre
M Benz - Werk Bauen und Wohnen, 2011
[Cite](#) [Save](#)

Se desconoce el origen de los metadatos. La referencia es correcta.

[CITATION] **Den Bauablauf gemeinsam steuern-Gute Zusammenarbeit zwischen Planern und Ausführenden**
R Haderstorfer - Garten und Landschaft, 2012
[Cite](#) [Save](#)

Se desconoce el origen de los metadatos. La referencia es correcta.

GRAN BRETAÑA

[CITATION] **Star City. The Future under Communism**
A Blood - 2010 - BURLINGTON MAG PUBL LTD 14- ...
[Cite](#) [Save](#) [More](#)

Aparece con un subtítulo diferente al de Avery. El título, autor, revista, año y paginación coinciden. Se desconoce el origen de los metadatos.

[CITATION] **Michelangelo's Laurentian Library: Drawings and Design Process**
JG Cooper - ARCHITECTURAL HISTORY, 2011 - SOC ARCHITECT HIST GREAT ...
[Cite](#) [Save](#) [More](#)

Referencia completa. ¿"Soc Architect Hist Great..." es un indicador del origen de los metadatos? No se ha encontrado ninguna pista clara,

[CITATION] **Into the Belly of the Beast: Exploring London's Victorian Sewers**
P Guillery - 2011 - ... WALK, LEEDS LS3 1AB, W YORKS, ...
[Cite](#) [Save](#)

En la referencia no aparece el título de la revista. Posteriormente aparece una ¿direccion? El artículo no se recupera a través del buscador general de Google.

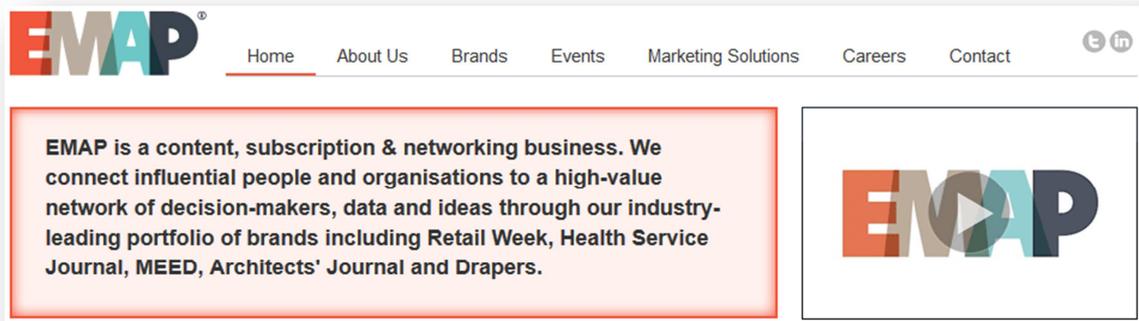
[CITATION] **LORE UNTO HIMSELF**-Meet Simon Costin, a fashion set designer and a magus of myths. His Hackney home, mixing march hares with Daleks, proves ...
S Patience - World of Interiors, 2010
[Cite](#) [Save](#)

Referencia completa. Se desconoce el origen del registro.

[CITATION] **Poundbury, UK**
A Powers - 2013 - EMAP BUSINESS PUBLISHING LTD ...
[Cite](#) [Save](#)

Referencia incompleta, no se indica el título de la revista "Architectural review". En cambio, ¿"EMAP Business Publishing LTD" puede ser una pista del origen de los

metadatos? Se ha buscado por la expresión en Google y se ha encontrado la página web de EMAP. En la página principal indican el objeto de la empresa, que consiste en ofrecer un servicio a las empresas de información y suscripción a bases de datos y revistas. Se centran en varias áreas entre las que se encuentra la Arquitectura.



Para el sector de la Arquitectura, trabajan en concreto con dos revistas: The Architectural Review (AR), al que pertenece el artículo del registro, y Architects Journal (AJ), <http://www.emap.com/products>.

