Búsqueda federada en el ecosistema de la e-ciencia: el caso *Science Research*

Por **Lluís Codina, Ernest Abadal** y **Cristòfol Rovira**

Resumen: *Se analiza el buscador Science Research en el contexto de la e-ciencia y de la búsqueda federada en comparación con la indización y la recolección. Se pone en relación también con otros buscadores académicos, especialmente Scirus y Google Scholar. Se realiza un análisis de los diversos componentes de Science Research y un estudio comparativo de obtención de resultados.*

Palabras clave: *Science Research, Deep Web, Buscadores académicos, Scirus, Google Scholar.*

Title: Federated search in the e-science ecosystem: *Science Research* case study

Abstract: *Analysis of the Science Research system in the context of e-science and the federated search, compared with the index-based retrieval systems and harvesting systems. Analysis of several elements of Science Research and a comparative study of search results.*

Lluís Codina *es profesor titular del Departamento de Comunicación de la Universidad Pompeu Fabra (UPF) y director de la Unidad de Soporte a la Calidad y a la Innovación Docente (Usquid) de la Facultad de Comunicación de la UPF. Imparte docencia en las titulaciones de Periodismo y de Comunicación Audiovisual. Es coordinador del Grupo de Investigación DigiDoc de la UPF y codirector del Máster Online en Buscadores y del Máster Online en Documentación Digital. Participa en el Máster Interuniversitario UB/UPF sobre Gestión de Contenidos Digitales.*

Ernest Abadal*, licenciado en filosofía, diplomado en biblioteconomía y documentación, y doctor en ciencias de la información, es profesor titular de la Facultad de Biblioteconomía y Documentación de la Universitat de Barcelona. Co-director del grupo de investigación "Acceso abierto a la ciencia". Autor de varios libros y artículos sobre publicaciones digitales y sobre la aplicación de las tecnologías de la información a la gestión de documentos. Director de la revista digital "BiD: textos universitaris de biblioteconomia i documentació".*

Cristòfol Rovira *es profesor del área de Biblioteconomía y Documentación de la Universidad Pompeu Fabra. Imparte docencia en las titulaciones de Publicidad y Relaciones Públicas, Comunicación Audiovisual, así como en el Máster Online en Documentación Digital, Máster en Buscadores y en el Máster Interuniversitario UB/UPF sobre Gestión de Contenidos Digitales. Sus líneas de investigación se centran en nuevas herramientas para la evaluación automática de sedes web (DigiDocSpider). Forma parte del grupo de investigación DigiDoc del Departamento de Comunicación de la UPF.*

Keywords: *Science Research, Deep Web, Academic search engines, Scirus, Google Scholar.*

Codina, Lluís; **Abadal, Ernest**; **Rovira, Cristòfol**. "Búsqueda federada en el ecosistema de la e-ciencia: el caso *Science Research*". *El profesional de la información*, 2010, enero-febrero, v. 19, n. 1, pp. 77-85.

DOI: 10.3145/epi.2010.ene.11

1. ¿Ciencia 2.0, 3.0?: e-ciencia

EXISTE UN AMPLIO CONSENSO EN CONSIDERAR la así llamada web 2.0 como responsable de los mayores cambios que han tenido lugar en internet desde mediados de la primera década del siglo XXI, y seguramente es la responsable de la imparable popularidad de la *World Wide Web* en relación con otros medios, canales o sistemas de comunicación.

En este sentido, parecía que solamente era cuestión de tiempo que estos cambios afectaran a las actividades académicas y de investigación.

En general hay dos ideas básicas subyacentes en la extrapolación de la web 2.0 al terreno de la ciencia: (1) la ciencia es comunicación; y (2) la ciencia es colaboración. Parece evidente que ambas cosas pueden mejorar con el uso de instrumentos como las redes sociales.

Esta idea está muy bien expresada por los fundadores de la red social para académicos *ResearchGate*: "The vision of science 2.0 is promising: communication between scientists will accelerate the distribution of new knowledge. [...] Science is collaboration, so scientific social networks will facilitate and improve the way scientists collaborate. Cooperation on scientific publications can be facilitated through wiki-like concepts" (*ResearchGate*, 2009).

No es extraño por tanto que desde hace ya algún tiempo se esté hablando de una ciencia 2.0 constituida por la aplicación de elementos de la web 2.0, concretamente las redes sociales y el *cloud computing*, a las tareas propias de la actividad académica e investigadora.

"Habría que hablar de una ciencia 3.0 ante la aparición de sistemas como el que nos ocupa"

También se habla de una web 3.0, una de cuyas características sería la relación entre aplicaciones diferentes, incluso heterogéneas para, entre otras cosas, combinar datos de procedencia distinta y presentarlos unificados en alguna forma que aporten un valor muy superior al de su existencia como tales datos o informaciones separados.

Un ejemplo sería el sistema de información científica que nos ocupa en este trabajo, el motor de búsquedas federadas producido por la empresa *Deep Web Technologies* denominado *Science Research*.

Si seguimos el juego de las denominaciones mediante números, habría que hablar tal vez de una ciencia 3.0 ante la aparición de sistemas como el aquí tratado. Dado que una supuesta ciencia 2.0 ó 3.0 tendría muy diversos componentes, todos ellos vinculados de una forma u otra con la *World Wide Web*, tal vez sería más conveniente adoptar la denominación más genérica de e-ciencia, en este caso como adaptación de la expresión original *e-science* procedente de la contracción de *electronic science*.

El inconveniente es que, como tantas otras veces, resulta una expresión forzada para el castellano. Pero este es el precio a pagar cuando el inglés no sólo se ha convertido en la principal lengua unificadora de la ciencia, sino que parece que además todas las innovaciones proceden del área cultural anglosajona. En realidad ya nos ha sucedido otras veces. Hoy nadie se sorprende del uso de la expresión I+D por ejemplo; por no mencionar cuando lo que se impone es el término original inglés: *know-how* o la propia palabra *web* (en lugar de telaraña) para referirse a una parte de internet.

La ventaja es que resulta una expresión que no necesitará actualizarse a nuevos dígitos (4.0, 5.0) cada vez que haya algún cambio tecnológico de una cierta entidad. El inconveniente es que la expresión e-ciencia en su origen se refería de forma muy concreta a la utilización de súper-ordenadores para la resolución de problemas que requieren cálculos masivos, por ejemplo simulaciones sobre la evolución del clima.

En todo caso, la cuestión que nos interesa aquí es recordar el nuevo contexto de la e-ciencia donde podemos situar una iniciativa como la del sistema de información *Science Research*; ecosistema en el que comparte el mismo hábitat pero tal vez no el mismo nicho que otras dos clases de soluciones a las que nos referiremos más adelante: los indizadores como *Scholar* o *Scirus* y los sistemas de las bibliotecas universitarias que permiten consultar de manera federada las colecciones suscritas por las mismas.

2. Indizar, recolectar, federar

Históricamente se han adoptado tres grandes tipos de soluciones tecnológicas a la hora de implementar sistemas de información capaces de vérselas con los contenidos de la Web: indización, recolección y federación.

– Indización: es la solución más antigua. Es lo que hacen la mayoría de los motores de búsqueda. Consiste, visto a 10 mil metros de altura, en lanzar un robot a identificar y copiar en su base de datos centralizada los contenidos más o menos dispersos en diferentes servidores, analizar ese contenido buscando cadenas de caracteres (palabras) y como resultado generar un índice. Cuando el internauta introduce una palabra clave, el motor la compara con las de su índice y en respuesta presenta una página de resultados con las referencias que coinciden.

– Recolección: fue la segunda solución en aparecer. En lugar de tener un complejo sistema de análisis e indización de documentos, la recolección (*harvesting*) es la creación de un índice común recogiendo metadatos de diversas fuentes que han sido codificados en forma de registros siguiendo normas y protocolos comunes. Uno de los más utilizados es el *OAI-PMH* (*Open Archives Initiative-Protocol for Metadata Harvesting*), creado en 2001 por la *Open Archives Initiative*. Un sistema de información tipo *harvesting* únicamente necesita mantener el índice común creado mediante su recolección, mientras que los registros y los documentos permanecen en las colecciones originales. Buenos ejemplos de este sistema son *OAIster* (anexionado recientemente por *OCLC*), *Narcis* (Países Bajos), *Arrow* (Australia) o *Recolecta* (España).

http://www.openarchives.org/
http://www.oclc.org/oaister/
http://www.narcis.info/
http://search.arrow.edu.au/
http://www.recolecta.net/

– Búsqueda federada: ha sido la última en surgir en el contexto que nos ocupa, pero se venía utilizando desde hace tiempo también en el campo de los buscadores. Consiste en enviar la misma pregunta a diversos motores simultáneamente. Un clásico de estas soluciones es *Metacrawler*. A este tipo de motores se les denomina meta-

buscadores o multibuscadores. Las bibliotecas universitarias también han instalado aplicaciones que posibilitan la consulta de forma conjunta del catálogo de la biblioteca junto con las distintas bases de datos y portales de revistas que tienen suscritas. Uno de los productos más extendidos en España es *Metalib*, de la empresa *ExLibris*.

http://www.metacrawler.com/
http://www.exlibrisgroup.com/

"La búsqueda federada consiste en enviar la misma pregunta a diversos motores"

Ahora bien, aplicar la búsqueda federada a una colección heterogénea de depósitos digitales, archivos, etc., obtener las respuestas y crear una página de resultados bien organizada requiere mucha más ingeniería que en el caso de los motores individuales, y aún más si se contempla el uso de búsqueda avanzada, que se expresa de una forma distinta en cada "colección", por usar la terminología de *Science Research*.

Como parece evidente, cada uno de los tres sistemas descritos presenta un balance de ventajas e inconvenientes, por ello todos conviven en este momento en el ecosistema de la e-ciencia y probablemente lo seguirán haciendo en el futuro. La tabla 1 constituye un resumen de lo que hemos comentado hasta ahora y añade la perspectiva del balance de ventajas e inconvenientes de cada solución.

3. *Science Research*

3.1. Contexto

Science Research es un producto de la empresa norteamericana *Deep Web Technologies*, fundada en el año 2002 por el ingeniero **Abe Lederman** aprovechando su larga experiencia de trabajo en el campo de las búsquedas federadas, los sistemas de gestión documental y la gestión del conocimiento en el seno de diversas empresas (*HP*, *Verity*) y como consultor de varias agencias gubernamentales vinculadas con ciencia y tecnología.

Solución	Ventajas	Inconvenientes	Ejemplos en e-ciencia
Indizar	– Respuestas muy rápidas. – Altas posibilidades de análisis de la información (minería de datos). – Alta calidad en la ordenación y presentación de los resultados. – No requiere acuerdos previos con terceros ni generar protocolos comunes.	– Alta exigencia y altos requerimientos en recursos de computación. – Desfase entre la publicación de contenidos y su tratamiento.	*Scirus* *http://www.scirus.com*
Recolectar	– Posibilidad de utilización de sistemas de metadatos sofisticados. – Escaso ruido y respuestas muy precisas. – Resultados muy bien estructurados. – Menores exigencias en el procesamiento de la información.	– No siempre se indiza el texto completo. Menores posibilidades de análisis y de recuperación de la información. – Requiere la elaboración de protocolos y acuerdos formales con terceros.	*OAIster* *http://www.oaister.org* *Arrow* (*Australian Research Repositories Online to the World*) *http://arrow.edu.au* *Recolecta* *http://recolecta.net*
Federar	– No requiere descargar, copiar, almacenar ni analizar documentos. – No hay desfase entre la publicación de contenidos y la incorporación al sistema de búsqueda. – Posibilidad de rigurosa selección de las fuentes. – No son imprescindibles acuerdos formales ni protocolos comunes.	– Lenguaje de consulta limitado al mínimo común. – Posibilidad de alta tasa de resultados duplicados. – Tiempos de respuesta más lentos. – Exige ingeniería en el envío de la misma consulta a centenares de fuentes diversas y en la compilación y presentación de resultados.	*Science Research* *http://www.scienceresearch.com*

Tabla 1. Comparación de las tres tecnologías básicas de los sistemas de información heterogéneos

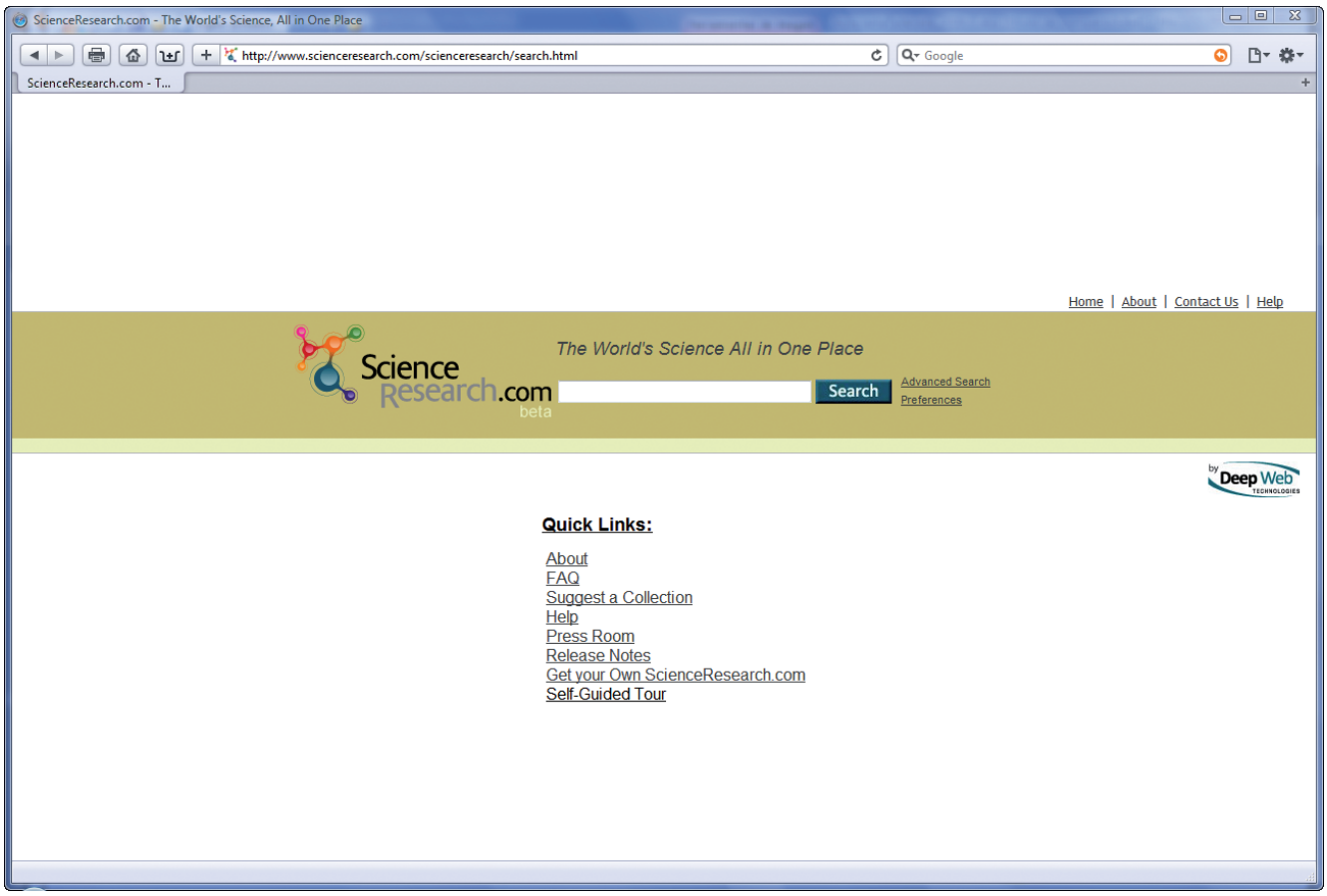

Figura 1. Página principal de Science Research con la interfaz de consulta muy simple, obviamente inspirada en Google
http://www.scienceresearch.com

En el mercado hay no obstante otros servicios de acceso a información científica que utilizan la misma tecnología (*DeepWeb*) y que tienen una estructura y funcionamiento similares a *ScienceResearch*. Se trata de los siguientes:

– *Biznar*: creado directamente por *DeepWeb* en 2008. Se centra en el campo de la información económica y de negocios.

http://www.biznar.com

– *Scitopia*: creado por el *American Institute of Physics* en 2007. Facilita la consulta de las colecciones digitales de veintiuna sociedades científicas y asociaciones técnicas y profesionales, entre las que destaca el *IEEE*.

http://www.scitopia.org

– *Worldwidescience*: creado por la *Oficina de Ciencia* del *Departamento de Energía* de los EUA también en 2007. Facilita la consulta de 44 grandes bases de datos de ciencia de 32 países, entre las que se encuentra, por ejemplo, el *Inist* francés o el portal *Scielo*, que incluye contenidos españoles.

http://worldwidescience.org

La empresa *Deep Web Technologies*, según indica en su web, aporta además soluciones tecnológicas en el campo de la gestión documental a organizaciones gubernamentales y a diversas empresas *Fortune 500*[1].

Por tanto, parece fácil deducir que el papel de *Science Research* (y de los otros productos antes señalados) en el modelo de negocio de la empresa es servir de escaparate y de promoción de las soluciones tecnológicas de la misma.

Google ya demostró en su momento la viabilidad de esta fórmula, al menos antes de que descubrieran el negocio de los anuncios. Además, si *Science Research* tiene éxito en todos los sentidos, es decir, si tiene un gran número de usuarios y si cada vez más colecciones confían en su fórmula tecnológica, parece también fácil deducir que *Science Research* es la plataforma de análisis y de estudios que cualquier empresa de su campo soñaría. Si además beneficia a los ciudadanos en general aportando un buen sistema abierto de información para académicos, la verdad es que sólo podemos felicitarnos.

No estamos diciendo que el sistema sea perfecto, en los siguientes apartados presentaremos nuestras críticas. Estamos diciendo que la fórmula parece ideal: una empresa quiere demostrar su poderío tecnológico poniéndolo al servicio de todos de forma abierta.

3.2. Componentes y funciones

Ya hemos señalado que *Science Research* (*SR* a partir de ahora) utiliza la búsqueda federada. Sus componentes son las colecciones a las que envía las preguntas, a partir

de las cuales compila las respuestas y las organiza en su página de resultados.

Colecciones

SR denomina así a las casi 400 fuentes que utiliza. Teniendo en cuenta la dimensión, la calidad, pero también la heterogeneidad de las colecciones, lo cierto es que, con todos sus fallos, es una proeza tecnológica.

La cuestión es que algunas de estas colecciones son a su vez colecciones de colecciones, y otras son motores de búsqueda. Por este motivo la cantidad total de información a la que podemos acceder mediante *SR* es virtualmente ilimitada, pero también muy redundante. Podemos ver esto con mayor claridad si examinamos unos cuantos ejemplos concretos.

"La página de resultados los categoriza por cinco criterios: temas, autores, publicación, editores y fechas"

La lista de las casi 400 fuentes incluye colecciones de asociaciones científicas y profesionales, como por ejemplo:

– *American Society for Biochemistry and Molecular Biology*

– *Association for Computing Machinery (ACM)*

– *Institute of Electrical & Electronics Engineers (IEEE)*

– *NASA Technical Reports Server*

Pero también repertorios que incluyen a su vez documentos procedentes de asociaciones como las anteriores, entre otros componentes, como por ejemplo:

– *BioMed Central*

– *Directory of Open Access Journals (DOAJ)*

– *Intute*

– *OAIster*

Editoriales y revistas científicas:

– *HighWire Press*

– *IngentaConnect*

– *National Academies Press*

– *Nature Publishing Group*

Finalmente, aunque esto no agota la tipología, bases de datos de patentes y motores de búsqueda:

– *Google Scholar*

– *European Patents*

– *US Patent and Trademark Office*

Una lista de fuentes o colecciones como la anterior viene con dos noticias bajo el brazo, una buena y una mala. La buena es que parece que nada se va a escapar del alcance de *SR*. La mala es que la probabilidad de que casi cualquier documento aparezca en dos o más de las colecciones es muy elevada. Si el sistema fuera eficaz para detectar y eliminar duplicados esto no sería un problema, pero lo cierto es que los resultados están plagados de duplicados. Se puede considerar un inconveniente o el precio inevitable a pagar, pero ahí está.

Opciones de búsqueda y página de resultados

Búsqueda avanzada

El problema de la búsqueda federada es que las opciones deben limitarse a lo que es el mínimo común de todas las colecciones. Es decir, puede que una colección tenga opciones de búsqueda muy sofisticadas, pero cuando además hay que enviar la misma pregunta a otras muchas fuentes tales opciones no se pueden aprovechar.

En concreto, como podemos ver en la figura 2, las opciones consisten en buscar por (1) texto completo, (2) título, (3) autor, (4) rangos de fechas y (5) tipo de colección.

No se puede objetar mucho por los motivos ya señalados, pues es difícil poder enviar una misma pregunta simple a tantas fuentes heterogéneas. Cualquier opción adicional, por sencilla que sea, es una proeza técnica.

"*SR* es un ejemplo reciente de la consolidación de las búsquedas federadas"

Las opciones de *Google Scholar* son las mismas, pero añade la de buscar por publicación, cosa que puede ser realmente útil; en cambio no permite limitar la búsqueda a una colección o tema. Por su parte *Scirus* es el sistema que presenta la búsqueda avanzada más completa porque cuenta con las opciones que tiene *SR* pero no *Scholar*, con las que tiene *Scholar* pero no *SR* y añade aún la de tipo de documento (por ejemplo artículos, patentes, tesis, etc.), además de contar con el único lenguaje de búsqueda que posibilita el uso de máscaras y comodines para sustituir caracteres o grupos de caracteres. En la figura 2 podemos ver la búsqueda avanzada de *Scirus* (con algunas opciones desplegadas sólo parcialmente).

Página de resultados

Es interesante, principalmente por el componente que denomina "topics" que consiste en una categorización/distribución de resultados por cinco criterios distintos: (1) temas, (2) autores, (3) publicación, (4) editores y (5) fechas.

La página contiene los siguientes componentes: (1) una serie de posibles acciones a realizar con la lista de referencias, siendo tal vez la más destacable la que permite

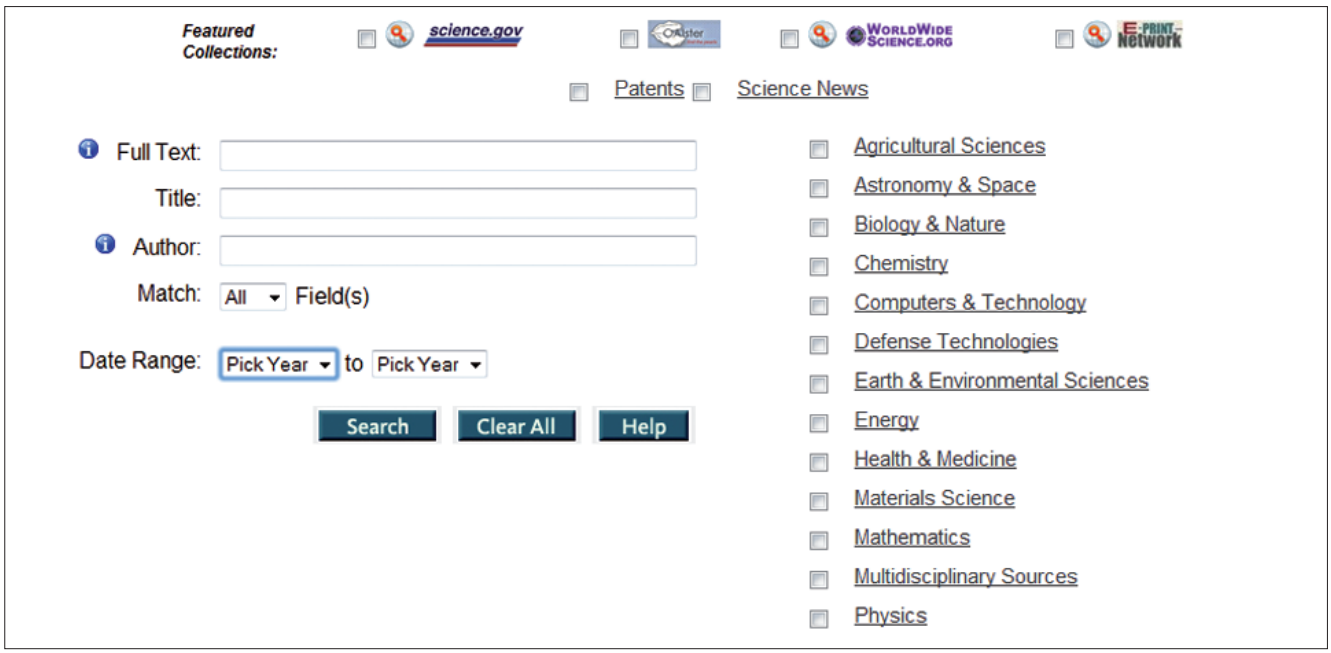

Figura 2. Página de búsqueda avanzada

crear listas de resultados seleccionados y exportarlos a *RefWorks*; (2) datos estadísticos; (3) información de estatus; (4) categorización y distribución de resultados según los criterios que hemos destacado antes; (5) opciones de navegación, ordenación y filtrado, posiblemente uno de los grupos de funciones más útiles; (6) ocupando la parte principal tenemos la lista en sí misma.

Desde un punto de vista pragmático y funcional, la página muestra al menos dos problemas, siempre si la comparamos con otros sistemas, principalmente con *Google Scholar* y con *Scirus*.

El primero es que, como ya hemos señalado, está repleta de resultados duplicados (de hecho, multiplicados: en algún caso hasta cuatro veces en la misma página).

El segundo es que la descripción no es homogénea. Después de un buen rato de trabajar con *SR* se echa de menos la predictibilidad típica de los sistemas basados en indizar o en recolectar, donde cada ítem de la página está descrito de forma sistemática.

En el caso de *SR* la descripción siempre está articulada en tres apar-

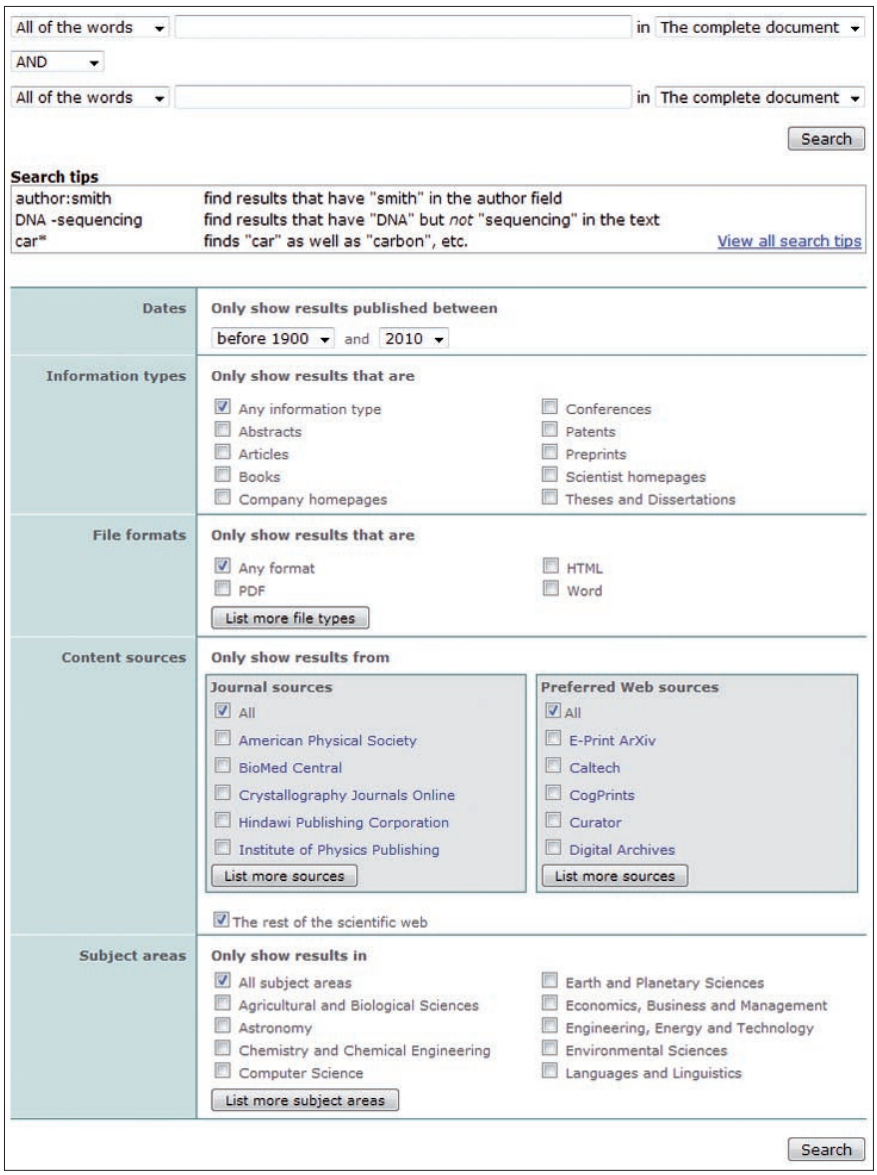

Figura 3. Búsqueda avanzada de Scirus

Figura 4. Lista de resultados

"Los buscadores como *Google Scholar* o *Scirus*, se basan en la indización, mientras que *Science Research* se basa en la búsqueda federada"

tados, como si fuera un bocadillo: en la parte superior, el título; en la parte inferior el nombre de la colección de procedencia y en el medio, una información complementaria y el resumen del documento. El problema es que el resumen de la parte central, a veces no existe; otras veces son unas líneas de texto poco inteligibles. Solamente en algunas ocasiones consiste en lo que se supone que debería ser siempre: hasta dos líneas que actúan como resumen. Por lo que hace a la información complementaria que forma también el centro del bocadillo imaginario a veces es el nombre del autor, a veces el nombre de la publicación o de la fuente; esta última también se encuentra de formas diversas. Además aparece una calificación basada en estrellas que no está documentada en la ayuda oficial.

Creemos que son demasiadas variaciones y una gran falta de coherencia para que el usuario se sienta seguro utilizando este sistema. Al menos ésta es la sensación que tuvimos en la realización de las pruebas. La figura 5 destaca esta falta de homogeneidad. Los resultados 4 y 5 carecen de resumen; 1 y 3 indican el autor, mientras que 2 y 6 indican la publicación.

Por último, se dispone de la opción denominada *collection status* que permite saber qué colecciones han aportado resultados a la página, cuántos en total y cuántos de los mismos han sido añadidos al resultado global. La figura 6 muestra una vista parcial de esta lista.

3.3. Estudio cuantitativo

Además de las pruebas y análisis funcionales que ya hemos comentado, deseábamos saber el número de resultados que es capaz de proporcionar *SR* y compararlos con los de sus dos mejores rivales, tal como hemos venido haciendo hasta ahora: *Google Scholar* y *Scirus*. La tabla 2 muestra los resultados de lanzar las mismas 10 búsquedas a los tres motores.

Todas las búsquedas se llevaron a cabo el mismo día en la tercera semana de julio de 2009. Para todas las preguntas se usaron comillas y alfabeto pobre (por ejemplo: "web semantica" y no Web Semántica).

En la columna de *SR* vemos que aparecen dos resultados. El primero, que siempre es menor, es el número de resultados recolectados por *SR*. El segundo, entre paréntesis, es el número total identificado. La primera cifra es la única operativa porque es el número de los resultados a los que podemos acceder.

Por ejemplo, si para la pregunta n. 1 vemos el resultado 765 (311.071), esto significa que, aunque *SR* identifica en el total de las colecciones más de 300 mil resultados (y así nos lo dice), ha compilado para nosotros un total de 765 y por tanto no se puede acceder a los

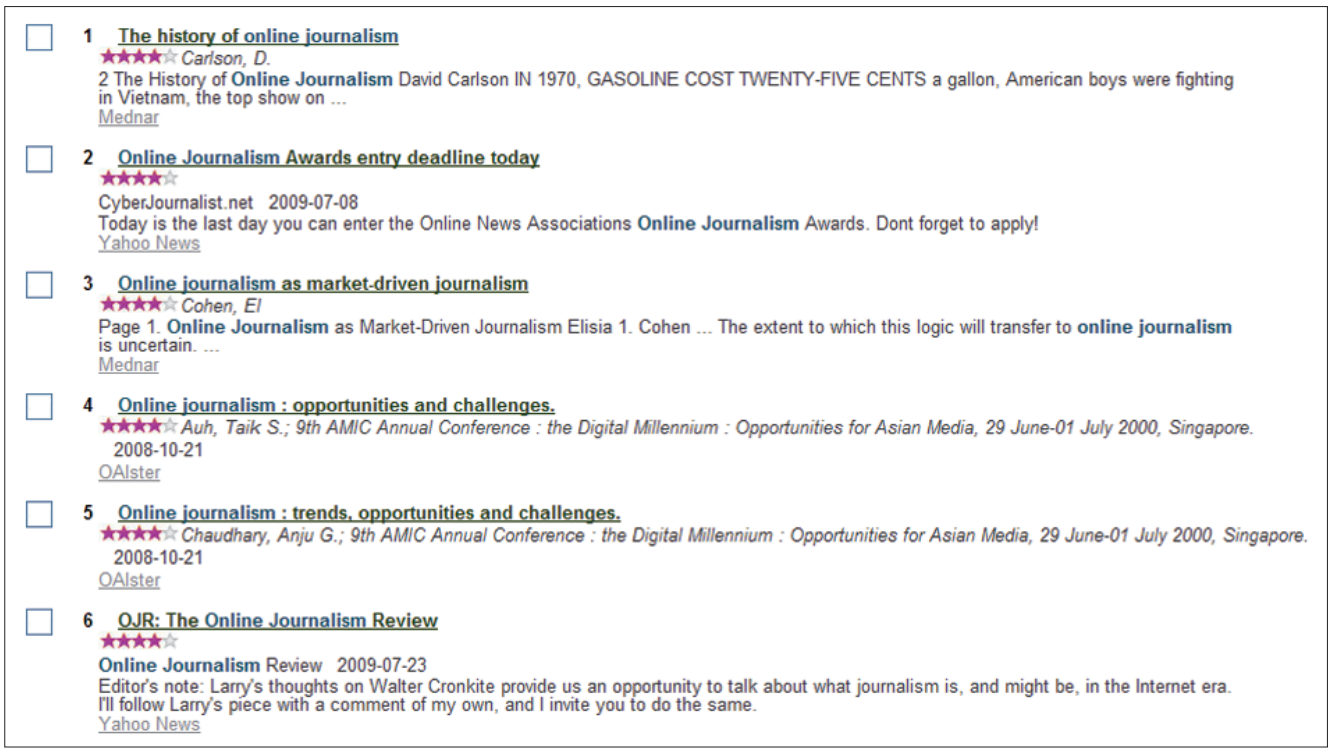

Figura 5. Página de resultados

documentos a partir del resultado número 766. ¿Por qué no hay siempre el mismo número máximo, por ejemplo, siempre 1.000 o siempre 5.000? No lo sabemos. No hemos encontrado información en la ayuda oficial. ¿Por qué hay tanta variación en el ratio mostrados/encontrados? La respuesta en este caso es la misma, no hay explicación en la documentación oficial del sitio. Para *Scirus* y *Google Scholar* no surgen tantos interrogantes puesto que en todas las búsquedas se compilan siempre como máximo 1.000 resultados.

Si nos atenemos a las cifras efectivas (la primera que aparece en el caso de *SR*) en cuanto a números absolutos vemos que los de *Scirus* y *Scholar* son muy inferiores cuantitativamente hablando, en algunos casos incluso en varios órdenes de magnitud (véase las búsquedas número 2, 8 y 10).

Otra cuestión sería determinar la relevancia de los primeros resultados, digamos de los 10 ó 20 primeros ya que ante cifras de varios miles y superiores lo único que de verdad importa es el acierto en el cálculo de relevancia. Determinar este aspecto requeriría realizar estudios de usuarios, cosa que escapa a los objetivos de este trabajo.

Por último, en las pruebas realizadas vimos que puede darse hasta un 50% de referencias duplicadas en la primera página de resultados; es decir, de las 10 referencias de la primera página, so-

n	Pregunta	*Science Research*	*Scirus*	*Google Scholar*
1	Web semántica	765 (311.071)	1.000 (3.390)	1.000 (5.670)
2	Semantic web	2.846 (3.722.306)	1.000 (343.910)	1.000 (106.000)
3	Web social	2.233 (325.191)	1.000 (8.087)	1.000 (1.790)
4	Social web	2.295 (318.478)	1.000 (81.968)	1.000 (4.530)
5	Periodismo digital	954 (14.291)	1.000 (2.348)	1.000 (1.050)
6	Online journalism	1.548 (31.755)	1.000 (49.056)	1.000 (3.310)
7	Cambio climático	1.163 (2.526)	1.000 (85.453)	1.000 (22.200)
8	Climate change	5.755 (33.277.639)	1.000 (2.796.048)	1.000 (896.000)
9	Evolución humana	1.053 (1.450)	1.000 (3.551)	1.000 (4.580)
10	Human evolution	3.892 (679.773)	1.000 (257.814)	1.000 (130.000)

Tabla 2. Resultados de las mismas 10 búsquedas en los tres motores

ASM International	✓	0	0
Association for Computing Machinery	✓	20	20
Astronomical Journal, The	✓	0	0
Astronomy & Astrophysics	✓	6	6
Atmospheric Radiation Measurement Program	✓	0	0
AULIMP (Air University Library)	✓	20	195
Bandolier	✓	0	0
Bioenergy Feedstock Information Network	✓	0	0
BioMed Central	✓	14	14
BioOne	✓	0	0
Biophysical Journal	✓	0	0
Biotechnology Industry Organization	✓	0	0
Blekinge Institute of Technology: Electronic Research Archive	✓	0	0
British Library Direct	✓	9	13

Figura 6. Vista parcial de la lista de colecciones con los resultados obtenidos de cada una

lamente 5 eran referencias únicas, las otras 5 eran una suma de casos de (multi)duplicación. En concreto para la pregunta 04 hasta 3 documentos únicos presentaban casos de duplicación sumando 5 documentos duplicados: *Informe APEI sobre web social* (4 ocurrencias, por tanto, duplicado tres veces), *Las 10 claves de la web social* (2 ocurrencias, duplicado una vez) y *Rankkit: Web social de encuestas* (2 ocurrencias, duplicado una vez).

La pauta se repetía en las páginas siguientes. Ahora bien, estudios limitados a la primera página de resultados mostraron que la tasa de duplicados (para las preguntas de la muestra) podían llegar al 50%. En concreto, 6 de las 10 preguntas de nuestro estudio arrojaron duplicados, y en tal caso, las duplicaciones oscilaban entre el 20 y el 50%. Por ejemplo, en el peor de los casos (50%) de los 10 resultados, solamente 5 eran resultados únicos, los otros 5 eran casos en los que un mismo resultado aparecía al menos dos veces.

4. Conclusiones

SR es un ejemplo reciente de la consolidación de las búsquedas federadas y tal vez el inicio de una nueva generación de servicios de información científica basados en esta tecnología (por cierto, varios de ellos basados en la misma solución que *SR*, es decir en *DeepWeb*).

A diferencia de otras experiencias anteriores parece que ésta se basa en un número mucho más amplio de colecciones y ha querido adoptar el formato de buscador al que está acostumbrado un público muy amplio gracias a *Google*.

Esto significa que en estos momentos hay, por un lado, dos clases de sistemas que compiten por ofrecer soluciones parecidas: (1) los buscadores como *Google Scholar* o *Scirus*, basados en la indización, y (2) los buscadores independientes como *SR* y otros como los mencionados. Pero además, hay una tercera clase: los sistemas instalados por las bibliotecas universitarias para consultar de manera federada las colecciones suscritas. Para el usuario, particularmente si pensamos en un público universitario con acceso a la tercera clase de sistemas son, de hecho, tres soluciones que se solapan en parte entre ellas.

Parece que la solución de *SR*, al menos en su forma actual, aún no está del todo madura, de manera que muchos usuarios podrían seguir prefiriendo el uso de *Scirus* por ejemplo, o bien el metabuscador que le ofrezca su biblioteca (o ambos, por supuesto).

Naturalmente está por ver cómo evolucionará esta nueva forma de búsqueda federada. Puede que en el futuro se consolide y sea una tercera solución que encuentre su propio nicho frente a *Scirus* y *Google* (por ejemplo entre investigadores y estudiosos no universitarios). Por lo que hace al entorno universitario, habría que llevar a cabo otros análisis para constatar su utilidad frente a los sistemas de búsqueda federada de las bibliotecas universitarias.

En todo caso sólo cabe felicitarse de que haya nuevas iniciativas tecnológicas y empresariales en este terreno y por tanto, desearle la mejor suerte a *Science Research*. La buena competencia ya ha demostrado otras veces que mejora los productos y hace crecer el mercado.

5. Nota

1. Se conoce así a las 500 mayores empresas de EUA según el ranking de la revista *Fortune*.

6. Bibliografía

Boswell, Wendy. *Online research*. Avon: Adams Media, 2007.

Cabezas-Clavijo, Álvaro; **Torres-Salinas, Daniel**; **Delgado-López-Cózar, Emilio**. "Ciencia 2.0: catálogo de herramientas e implicaciones para la actividad investigadora". *El profesional de la información*, 2009, enero-febrero, v. 18, n. 1, pp. 72-79.

Cervone, Frank. "Federated searching: today, tomorrow and the future (?)". *Serials*, 2007, March, v. 20, n. 1, pp. 67-70.

Codina, Lluís. "Ciencia 2.0: redes sociales y aplicaciones en línea para académicos". *Hipertext.net*, 2009, n. 7.
http://www.hipertext.net/web/pag295.htm

Fingerman, Susan. "Scitopia: worthy effort; worth your effort?". *Online*, 2008, Jul.-Aug., v. 32, n. 4, pp. 28-31.

Krosky, Ellyssa. *Web 2.0 for librarians and information professionals*. New York: Neal-Schuman, 2008.

Ojala, Marydee. "Search engines designed for business". *Online*, 2009, Jan.-Feb., p. 44-46.

Lluís Codina, ***Cristòfol Rovira***, *Universitat Pompeu Fabra, Barcelona.*
lluis.codina@upf.edu
cristofol.rovira@upf.edu

Ernest Abadal, *Universitat de Barcelona.*
abadal@ub.edu