Sistemas de búsqueda y obtención de información: componentes y evolución

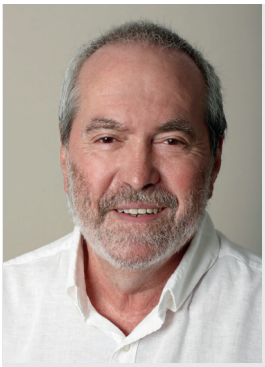
Information seeking and retrieval systems: Components and evolution

Lluís Codina

Codina, Lluís (2018). "Sistemas de búsqueda y obtención de información: componentes y evolución". *Anuario ThinkEPI*, v. 12, pp. 77-82.

https://doi.org/10.3145/thinkepi.2018.06

Publicado en *IweTel* el 29 de diciembre de 2017



Resumen: Presentación y descripción de los componentes universales de los sistemas de búsqueda y recuperación de información. Sistema de búsqueda, página de resultados y modelo de registro (metadatos). Consideración de los sistemas de información documental como entorno de los sistemas de búsqueda. Principales tendencias en la evolución reciente de sus interfaces de usuario en el caso de las bases de datos académicas.

Palabras clave: Sistemas de información documental; Búsqueda y recuperación de información; Interfaces de usuario; Bases de datos académicas.

Abstract: A description and presentation of the universal components of information seeking and retrieval systems including query system, search results page, and record (metadata) model. Bibliographic information systems as an environment for bibliographic retrieval systems is considered. The main trends in the recent evolution of user interfaces in the case of academic databases are discussed.

Keywords: Bibliographic information systems; Information seeking: Information retrieval; Information systems; User interfaces; Academic databases.

1. Introducción

Décadas de desarrollo de sistemas de información documental, con sus elementos correspondientes de recuperación de información, han dado lugar a un consenso (a veces más implícito que explícito) sobre cuáles son los componentes de tales sistemas y las formas más eficaces de implementarlos mediante sus interfaces de usuario (**Nielsen**; **Lorangen**, 2006; **Hearst**, 2009; **Thurow**; **Musica**, 2009; **Morville**; **Callender**, 2010; **Abadal**; **Codina**, 2011; **Guallar**; **Codina**; **Abadal**, 2017).

En este trabajo nos proponemos presentar un balance global de tales aspectos. En primer lugar, revisando las bases generales de la búsqueda de información, para pasar después a mostrar los componentes característicos de un sistema de información documental, y finalmente, mostrar los patrones o formas más habituales de implementar las funciones correspondientes.

Introduciremos también una pequeña revisión sobre las últimas tendencias en interfaces de búsqueda, tomando como referencia, principalmente, las interfaces y las funciones de búsqueda de bases de datos académicas como *Scopus* o *Web of Science* (**Borrego**; **Anglada**, 2016; **Codina**, 2017).

2. Búsqueda y obtención de información

La búsqueda de información es una de las actividades vitales del ser humano. No es una actividad vinculada exclusivamente con internet y ni siquiera es una actividad nueva (**Battelle**, 2006). Es tan antigua como la humanidad. Lo que marca la diferencia con nuestro tiempo es la súper abundancia de información, pero cuanta mayor oferta de información existe, más necesarios son los sistemas eficientes de búsqueda. Lo cierto es que, actualmente disponemos también de una oferta realmente exuberante de ellos, formando así un auténtico ecosistema, con diferentes nichos y relaciones entre los mismos.

Por este motivo, la disciplina que presentaremos a continuación tiene al menos dos facetas, cada una con diversos componentes:

- la más teórica se centra en el estudio de las características generales o abstractas de los sistemas de información, particularmente de tipo documental;
- la más aplicada se centra en el conocimiento de los sistemas de información existentes para cada rama del conocimiento o para cada tipo de necesidad de información.

La búsqueda y obtención de información es el nombre del correspondiente campo de estudios (*information seeking and retrieval* en inglés). Tiene muchos puntos en contacto con la recuperación de información (RI), pero es una disciplina independiente y con la que, por supuesto, se complementa en lo que podría consistir el campo más amplio de la búsqueda de información (**Hearst**; 2009; **Gardner**; **Inger**, 2016).

Se diferencia de la RI porque tiene un enfoque cognitivo y no (necesariamente) algorítmico. Se ocupa del estudio de los sistemas de información existentes, de la oferta de tales sistemas en el mundo real y de las características de sus interfaces de usuario más que de los modelos formales o abstractos que hay detrás de estos sistemas de información.

3. La búsqueda como sector económico y social

La búsqueda, además de ser una actividad y un campo de estudios, como estamos señalando, es también un sector económico (el cual a su vez está generando su propio campo de estudio, la economía de la búsqueda).

Una actividad que practica a diario más del noventa por ciento de los ciudadanos que poseen una conexión a internet, no podía escapar a la dimensión económica. El sector más conocido es el de los buscadores de página web como *Google* o *Bing*, pero hay otros actores, como los buscadores académicos o los bancos de fotografía y vídeo.

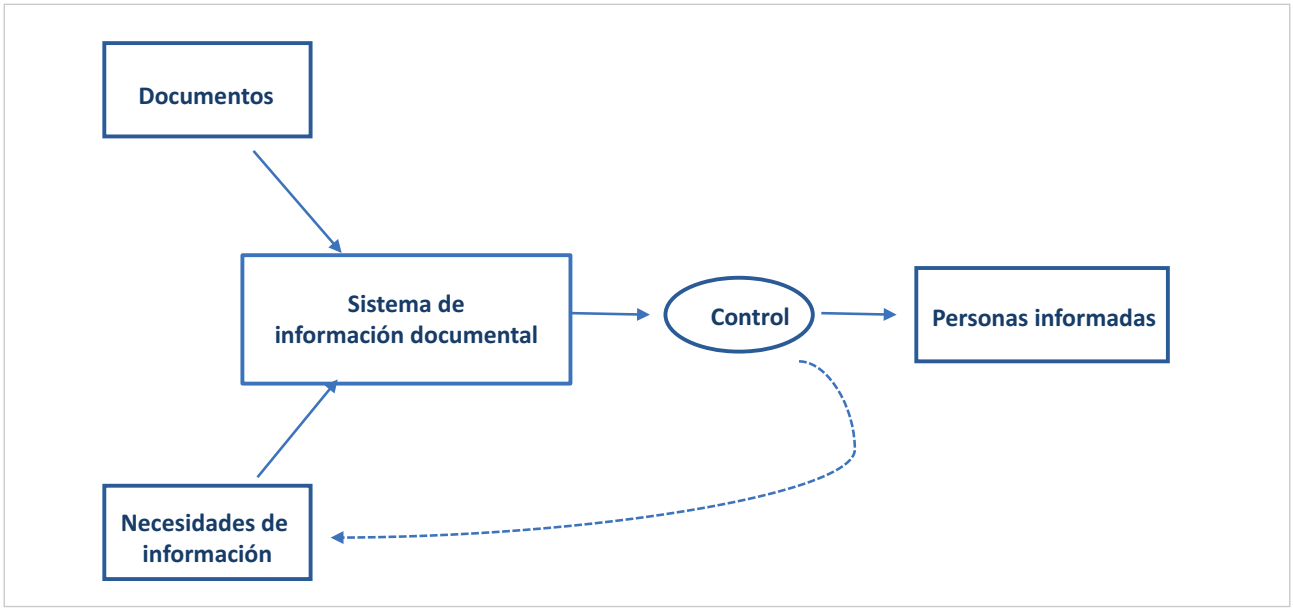


Imagen 1. Diagrama de un Sistema de información documental

4. Sistemas de información documental (SID)

Para relacionar lo anterior con las necesidades de información de los usuarios es conveniente concebir las funciones de un sistema de información documental (SID, a partir de ahora) como una clase de sistema que se puede representar con el diagrama que mostramos en la figura 1.

Podemos ver que un sistema documental (tal como un motor de búsqueda) acepta dos clases de entradas:

- documentos: en nuestro caso son páginas web principalmente;
- necesidades de información: las preguntas de los internautas.

La salida del sistema es una lista de sitios web que, presumiblemente, contienen información capaz de satisfacer las necesidades de información del usuario, de aquí que la salida se rotule con el (optimista) texto de Personas informadas.

Naturalmente, si la primera consulta no da buenos resultados, lo lógico es que el usuario (el internauta) haga algún intento más. Esto queda representado por la elipse rotulada con la palabra Control. Es decir, se supone que el usuario examina la relevancia de los resultados y si no le satisfacen, reitera la pregunta (usando otros términos).

"La búsqueda de información se ocupa del estudio de los sistemas de información existentes, de la oferta de tales sistemas en el mundo real y de las características de sus interfaces de usuario"

5. Evolución de los sistemas de consulta

Como otros sistemas de información, las interfaces de búsqueda han ido evolucionando en el curso de los años, pese a que, a la vez, se mantienen de forma muy estable sus componentes esenciales desde el punto de vista de su relación con los usuarios.

Hemos considerado los sistemas de información documental como sistemas, lo que nos ha permitido examinar sus entradas, su proceso de transformación, y sus salidas. Ahora, si los consideramos desde el punto de vista de su interfaz de usuario, podemos decir que todos los sistemas de información documental tienen algunos o todos estos elementos (**Hearst**, 2009; **Thurow**; **Musica**, 2009; **Morville**; **Callender**, 2010; **Guallar**; **Codina**; **Abadal**, 2017):

a) Búsqueda
- Simple
- Avanzada

b) Página de resultados con opciones de:
- Filtrado
- Ordenación
- Análisis
- Descubrimiento

c) Registro
- Simple
- Avanzado

d) Utilidades complementarias para la gestión de:
- Alertas
- Listados
- Búsquedas
- Bibliografías

La primera tendencia en la evolución de los elementos anteriores (no siempre presentes en todos los SID) ha consistido, notablemente, en un diseño progresivamente centrado en el usuario (UCD, por *user centered design*), lo que ha producido sistemas que son, a la vez, más complejos en muchas ocasiones, y sin embargo más fáciles de usar incluso por parte de usuarios noveles.

En los SID de mayor complejidad, como es el caso de las bases de datos más importantes (por ejemplo, *Scopus* y *Web of Science*) su uso y explotación adecuados sigue requiriendo de una notable curva de aprendizaje, aunque quedan lejos de las dificultades de las arcanas interfaces previas a la década actual.

La segunda tendencia notable es el aumento de prestaciones vinculadas con el concepto de usuario registrado. Es decir, cada vez más, esta clase de sistemas permiten añadir capas de funciones extras si los usuarios en lugar de hacer un uso anónimo de los mismos, se registran en el sistema y se identifican (es decir, inician sesión) para su uso.

En tercer lugar, otra tendencia notable consiste en la mejora de los sistemas de ordenación de la página de resultados, probablemente por la poderosa influencia de la búsqueda en internet. A las ordenaciones clásicas de tipo alfabético y cronológico, ahora se suman ordenaciones por cálculo de relevancia, por citaciones recibidas, por número de descargas, etc.

Por último, como cuarta tendencia, podemos citar la fuerte influencia de la web 2.0 que se manifiesta en la conexión de todos o de buena parte de los SID con redes sociales y con otras aplicaciones. Dentro de esta tendencia, es frecuente que los resultados se puedan compartir con redes sociales desde la ficha o registro de cada referencia, incluso que algunas de las métricas (altmétricas) que manejan estos sistemas de información tengan relación con tales redes sociales.

Un segundo aspecto de esta tendencia, consiste en que muchos de los SID pueden conectarse con otras aplicaciones de los usuarios, típicamente con gestores bibliográficos personales como *Mendeley*, *Zotero* o *EndNote*.

Por último, siempre en esta misma tendencia, podemos observar cómo algunos SID adquieren (o tienden a adquirir) rasgos de red social, promoviendo de alguna forma el intercambio de información y la colaboración entre sus usuarios registrados.

Estas cuatro tendencias, se pueden dar de forma conjunta o parcial en cada tipo de SID, siendo las bases de datos, como veremos, el SID que, actualmente, es capaz de aglutinar la mayoría de ellas.

6. Búsqueda avanzada

Por su parte, si examinamos ahora el componente más interesante y complejo de la interfaz de búsqueda, ha ido quedando bien establecido que un buen formulario de búsqueda avanzada, debe satisfacer los siguientes requerimientos (**Hearst**, 2009; **Thurow**; **Musica**, 2009; **Morville**; **Callender**, 2010):

1. No ser la primera opción que se muestra al usuario: el motivo es que la búsqueda avanzada debe reservarse a usuarios expertos o, al menos, motivados. Para usuarios inexpertos (o no motivados), es intimidatoria, por un lado; y una constante fuente de errores por otro debido a la posible (y frecuente) mala comprensión de las opciones propias de la búsqueda avanzada.

2. Incluir diferentes herramientas, las más habituales suelen ser:

a) Búsquedas parametrizadas: es decir, búsquedas por campos, por secciones o categorías, por tipo de documento, etc.

b) Búsquedas con operadores lógicos: operadores AND, OR, NOT (ver más adelante el apartado dedicado a los operadores booleanos).

c) Búsqueda por frase: aplica un AND, reforzado con la exigencia de que no haya ninguna palabra entre los términos de búsqueda y que el orden sea el mismo que la frase de búsqueda. Suele utilizarse la convención de entrecomillar los términos de la frase. No obstante, se considera una buena práctica presentar la opción de forma explícita en el formulario

d) Búsquedas de proximidad: suele utilizarse el operador NEAR entre dos términos de búsqueda para especificar el número máximo de palabras entre ambos términos.

e) Búsquedas combinadas: booleanas + parámetricas, por ejemplo.

f) Búsqueda por índices: autores, títulos, etc.

Opcionalmente, algunos sistemas permiten la búsqueda por tesauros o algún otro tipo de lenguaje documental: clasificaciones, taxonomías, etc. En todo caso, cabe tener presente que no todas las búsquedas avanzadas incluyen a la vez todas las opciones, en especial las opciones 2a) a 2d), sino que pueden incluir alguna combinación de dos o de tres de las opciones, típicamente, 2a) + 2b), o 2a) + 2b) + 2d), etc.

3. Opción de ayuda. Debe haber una página de ayuda con información, consejos y ejemplos de búsqueda.

7. Página de resultados

Esta página contiene la lista de los documentos o registros que satisfacen la ecuación de búsqueda. Los componentes de esta página pueden llegar a ser realmente extensos, sobre todo en las bases de datos académicas o bibliográficas, pero los más habituales son los siguientes:

a) Una fila superior donde aparece la palabra (o frase) de búsqueda y el número de resultados obtenido.

b) Sugerencias de búsqueda: palabras con otro deletreo, sinónimos, términos relacionados, etc.

c) La lista de los resultados en sí, con algunas indicaciones básicas para cada resultado, que suelen ser:

- título del documento;
- breve descripción del documento (entre una y tres líneas);
- propiedades del documento según el caso (tamaño, formato, fecha, etc.);
- enlace al documento completo, si es el caso;
- posibilidad de marcar todos o parte de los resultados para realizar acciones posteriores. Típicamente son las siguientes: imprimir, enviar por correo, exportar (lo veremos más adelante), recomendar y, por último, guardar en el espacio personal en sistemas que permiten la identificación de usuarios.

d) Opciones de ordenación: al menos por los siguientes criterios:

- título;
- fuente;
- autor;
- orden cronológico;
- grado de relevancia.

e) Opciones de exportación: generalmente, en el caso de bases de datos bibliográficas, exportación a formatos de citación e incluso exportación directa a otras aplicaciones o sistemas de información (p.e. editores de texto o bases de datos bibliográficas personales)

f) Historia de búsqueda con la posibilidad de combinar los conjuntos de resultados anteriores

mediante operadores booleanos.

g) Página de ayuda con sugerencias para búsquedas con cero resultados.

Como hemos señalado, a los componentes anteriores se puede unir aún una larga lista de opciones, por ejemplo, análisis de citaciones de los documentos recuperados, posibilidad de recibir avisos cada vez que se cite ese documento en el futuro, etc.

Tal vez uno de los mejores ejemplos de una página de resultados plena de opciones, pero integradas de una forma muy usable, y no intimidatoria, sea la de *Google*. También son excelentes ejemplos de páginas de resultados las de las bases de datos académicas como *Web of Science* o *Scopus* (**Halevi**; **Moed**; **Bar-Ilan**, 2017).

8. Un caso particular: búsquedas avanzadas en bancos de imágenes y vídeo

En el mundo anglosajón el sector de los bancos de imágenes y vídeo se conoce como *stock photography* o *stock footage* (algo así como "fotografía almacenada" y "tomas almacenadas"), términos que se oponen al de fotografía o vídeo de encargo.

El motivo es que la primera es la fotografía o el vídeo que se utiliza en lugar de encargar expresamente a un fotógrafo o a un realizador que tome las imágenes. Por ejemplo, si una publicación impresa española, digamos un semanario sobre viajes, necesita ilustrar un reportaje escrito sobre un país africano, tiene al menos dos opciones: desplazar a uno de sus fotógrafos o adquirir las imágenes directamente de un banco. Igualmente, si el productor de un film necesita un plano aéreo de alguna región de África, puede encargar la filmación o puede adquirir un clip de vídeo con la filmación de un banco de vídeos.

Por tanto, este sector de la búsqueda es posible porque los bancos de imágenes y vídeo orientados a un público profesional proporcionan acceso a enormes colecciones de fotografías e ilustraciones. Dos de los más importantes a nivel internacional son *Getty Image* y *Shutterstock*

En cualquier caso, además de los mencionados existen decenas de bancos de imágenes de mayor o menor tamaño dentro del denominado sector *Microstock*, regido por el *low cost*, pero con colecciones normalmente centradas en la fotografía (sin vídeo), con imágenes menos exclusivas y con un menor coste en general.

Las necesidades de búsqueda avanzada de esta clase de sistemas difieren bastante de las de los sistemas documentales convencionales. Aunque se utilizan también las búsquedas parametrizadas y las búsquedas booleanas, hay diferencias sustanciales que se presentan a continuación.

En primer lugar, la imagen en sí misma no es indizable, al menos no de forma conceptual, como sí lo es, en cambio, un documento convencional que incluya un apartado textual más o menos amplio. Ciertamente existen sistemas de indexación automática de imágenes que pueden derivar ciertas propiedades de la imagen de forma automática, como colores, formas dominantes, texturas, etc. Pero ningún sistema automático puede deducir (al menos hasta ahora) que una fotografía con un niño y la bandera de un país connota y, por tanto, puede indizarse con el término "nacionalismo".

Lo anterior nos indica ya una de las necesidades de búsqueda básicas que deben cubrir los formularios de búsqueda de un banco de imagen o vídeo, y que, al menos, son las siguientes:

- Elementos icónicos presentes en la imagen (p.e., niño, bandera).
- Elementos conceptuales, tanto denotados, como connotados (p.e. infancia, nacionalismo, etc.).

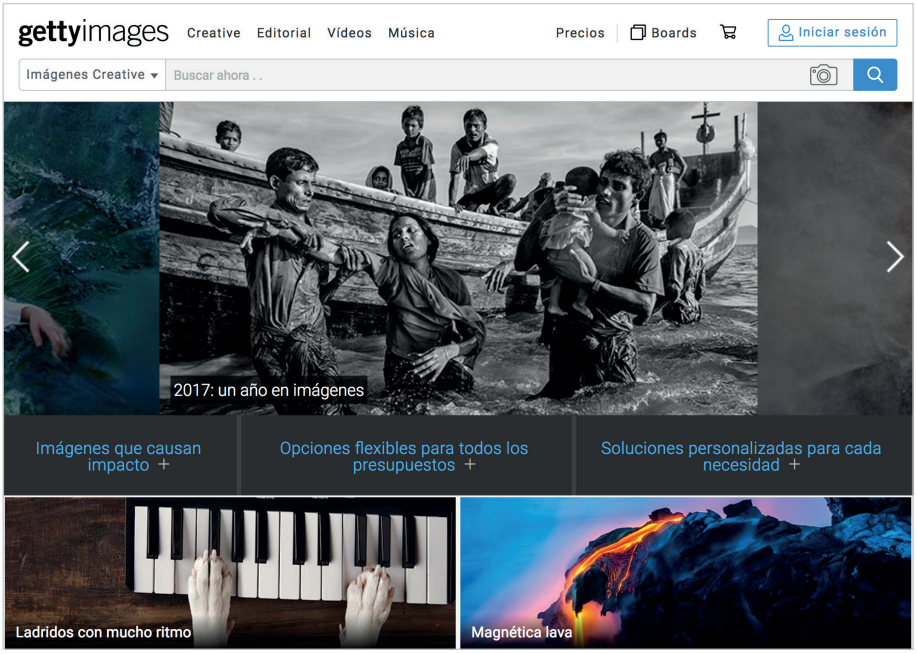


Imagen 1. *Gettyimages*
https://www.gettyimages.es

- Propiedades de la imagen (color, blanco y negro, etc.).
- Orientación de la imagen (vertical, horizontal, etc.).
- Características técnicas de la imagen (dimensiones, tamaño, etc.).
- Características legales (tipo de licencia, permiso del modelo, etc.).
- Características de la toma (lugar, fecha, interior, exterior, etc.).

Por último, dadas las necesidades de los usuarios de estos sistemas, la búsqueda avanzada suele proporcionar la siguiente pareja de posibilidades adicionales:

- especificar que no aparezcan personas en la imagen (suele utilizarse el descriptor "nadie");
- si hay personas, entonces poder especificar el número de personas que aparecen (suelen utilizarse los descriptores "una", "dos", "tres", "más de tres", "multitud").

"Cualquier sistema de información, sea del tipo que sea, debe estar orientado al usuario"

9. Conclusiones

Algunos analistas opinan que la búsqueda y obtención de información es el contexto imprescindible (por ser real, y no meramente de laboratorio) de la RI. Es de suponer que los practicantes de la RI desean que sus teorías y modelos tengan aplicación. Pues bien, tal aplicación, en caso de tener lugar sólo tiene sentido si se da en sistemas reales que utilizan usuarios reales, y no solamente en pruebas de laboratorio, por mucho que estas son también necesarias para el avance la disciplina.

Cualquier sistema de información, sea del tipo que sea, debe estar orientado al usuario. A estas alturas no hace falta insistir mucho en ello. No obstante, para dar satisfacción a esta exigencia son necesarios no solamente los estudios sobre algoritmos ideales, sino sobre sistemas de información realmente existentes, y que forman parte de sectores económicos y sociales muy importantes como los que hemos intentado representar aquí.

Nota

En el sitio web del autor, en la sección de análisis sobre bases de datos académicas, pueden encontrarse numerosos análisis de sistemas de búsqueda avanzada. *https://www.lluiscodina.com/cultura-academica/bases-de-datos*

10. Referencias

Abadal, Ernest; **Codina, Lluís** (2011). "El ecosistema de la información científica: estructura y niveles de agregación". *Anuario ThinkEPI*, v. 5, p. 128-131. *https://recyt.fecyt.es/index.php/ThinkEPI/article/view/30482*

Battelle, John (2006). *Buscar: cómo google y sus rivales han revolucionado los mercados y transformado nuestra cultura*. Barcelona: Urano. ISBN: 978 8493464233

Borrego, Ángel; **Anglada, Lluís** (2016). "Faculty information behaviour in the electronic environment: Attitudes towards searching, publishing and libraries". *New library world*, v. 117, n. 3/4, pp. 173–185. *https://doi.org/10.1108/NLW-11-2015-008*

Codina, Lluís (2017). "El ecosistema de la búsqueda académica: propuesta de caracterización". *Lluís Codina: comunicación y documentación*, 17 de julio. *https://www.lluiscodina.com/busqueda-academica-caracterizacion*

Gardner, Tracy; **Inger, Simon** (2016). *How readers discover content in scholarly publications: Trends in reader behaviour from 2005 to 2015*. Abingdon: Renew Training. ISBN: 978 0 9573920 4 5 *https://goo.gl/PhQk2Y*

Guallar, Javier; **Codina, Lluís**; **Abadal, Ernest** (2017) "Fuentes de información periodística". En: Gómez-Díaz Raquel; García-Rodríguez, Araceli; Cordón-García, José-Antonio. *Fuentes especializadas en Ciencias Sociales y Humanidades*. Madrid: Pirámide, p. 137-166. ISBN: 978 84 368 3790 2

Halevi, Gali; **Moed, Henk**; **Bar-Ilan, Judit** (2017). "Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the literature". *Journal of informetrics*, v. 11, n. 3, pp. 823–834. *https://doi.org/10.1016/j.joi.2017.06.005*

Hearst, Marti A. (2009). *Search user interface*. Cambridge: Cambridge UP. ISBN: 978 0521113793 *https://searchuserinterfaces.com/book*

Morville, Peter; **Callender, Jeffery** (2010). *Search patterns*. Sebastopol (CA): O'Reilly. ISBN: 978 0596802271

Nielsen, Jakob; **Loranger, Hoa** (2006). *Usabilidad: prioridad en el diseño web*. Madrid: Anaya. ISBN: 978 8441520929

Thurow, Shari; **Musica, Nick** (2009). *When search meets web usability*. Berkeley: New Riders. ISBN: 978 0321605894

Otras fuentes

Navigating Research: How academic users understand, discover, and utilize reference resources. Oxford: Oxford University Press, 2017. https://goo.gl/txHnuG

Wolff-Eisenberg, Christine; **Rod, Alisa**; **Schonfeld, Roger C.** (2016). *Ithaka S+R US faculty survey 2015*. New York: Ithaka S+R. *https://doi.org/10.18665/sr.277685*

Lluís Codina
Universitat Pompeu Fabra
lluis.codina@upf.edu