La WWW desde la perspectiva de la investigación en línea

Lluís Codina

El ciberespacio es cada vez más un lugar de publicación y de distribución de todo tipo de informaciones científicas y técnicas, perc también de todo tipo de informaciones sin más. Precisamente por ello, por la necesidad de «separar el grano de la paja», el ciberes ha dado lugar a un nuevo campo de actividad, la investigación en línea, lo cual podría constituir, por derecho propio, un nuevo can de estudios.

Cyberspace is increasingly becoming an area for publishing and distributing all types of scientific and technical information, as v lots of what we could simply term as information. Because of this need to separate the essentials from the non-essentials, cyberspa provided a new area of research, on-line research, which may become a new field deserving to be studied on its

La investigación en línea, desde nuestro punto de vista, estaría formada por cuatro componentes:

1. La búsqueda y obtención de información en la WWW (*World Wide Web*) en particular y en redes telemáticas en general. Éste sería, sin duda, el núcleo central de esta nueva disciplina. Incluiría también el estudio y la caracteriza de los hábitos de búsqueda de información de diversas comunidades de usuarios de la WWW con el fin de mejora interfases de usuario.

2. La evaluación y la descripción de recursos digitales, esto es, la determinación del valor de recursos tales como s webs, publicaciones digitales o bases de datos en línea y cómo organizar la descripción de sus propiedades de mar que sean reutilizables por parte de diversas comunidades de usuarios. Una parte de los esfuerzos en este campo se centran en los estudios de metadatos (debería hablarse en realidad de metainformaciones, ya que se trata de estudi cómo informar sobre la información) y normas, como el *Dublin Core*, relacionadas con los metadatos aplicados a comunidades del mundo de la ciencia, la cultura y la museística.

3. Los procedimientos para conseguir ser visibles en el ciberespacio sería un tercer y muy genuino componente de disciplina. Publicar en la WWW no significa que vayamos a ser vistos por un gran número de personas, sino más l todo lo contrario, a menos que se tomen algunas medidas especiales. Este componente, estrechamente relacionado el anterior, incluye el estudio de las mejores estrategias para representar los contenidos de una sede web o de cualc clase de recurso digital, de cara a obtener las mejores puntuaciones posibles (ránking) en los motores de búsqueda

4. Las características de la producción y la distribución de informaciones científicas a través de la WWW y qué cl de regularidades o patrones pueden derivarse de esa actividad. En este cuarto apartado tienen perfecta cabida la aplicación de las técnicas clásicas de la bibliometría y de la cienciometría (*scientometrics*) al ciberespacio.

Ignoramos si, en el futuro, estos cuatro aspectos que ahora, a algunos, se nos aparecen como elementos de una uni de estudio, se irán desagregando en el futuro o si, por el contrario, continuarán manteniéndose unidos. En el grupo investigación sobre representación del conocimiento, del Observatorio de la Comunicación Científica y del Institu Universitario de Lingüística Aplicada, ambos de la Universitat Pompeu Fabra, Barcelona, hacemos una apuesta pc consideración como parte de un mismo objeto de estudio y, en cualquier caso, nuestra propuesta de definición de l investigación en línea como campo de estudio es la siguiente: estudio de los aspectos de la WWW relacionados cc

producción, la difusión, la representación y la recuperación de recursos digitales de interés científico, cultural o técnico.

No examinaremos aquí, ni mucho menos, todos los aspectos mencionados, pero sí discutiremos algunas caracterís del ciberespacio relacionadas con la definición anterior.

Propiedades y características del ciberespacio

¿Cómo podríamos caracterizar el ciberespacio desde el punto de vista de la investigación en línea? Un retrato robo la WWW nos proporciona, de entrada, las siguientes características destacadas: es inmenso, es desorganizado y es desigual.

Dimensión

El ciberespacio es, efectivamente, inmenso, por lo menos de acuerdo con las dimensiones que solían tener los siste de información de la era pre-Internet.

Aunque existían ya grandes almacenes de información, como la famosa Biblioteca del Congreso o el distribuidor d base de datos Dialog, en toda la experiencia de la humanidad no ha habido ninguna clase de depósito, centro de documentación, biblioteca o sistema de información que creciera a la velocidad que lo hace Internet, de manera qu pronto contendrá (si no la contiene ya) la mayor masa de información unificada (es decir, en un mismo espacio) q haya producido nunca la humanidad.

En concreto, la WWW contenía, según estimaciones, entre 1400 y 2300 millones de documentos (páginas web) en noviembre de 1999 y creció a un ritmo del 4 % mensual durante 1999 (Dahn, 2000).

Ante tales dimensiones, la información en sí misma pierde valor y, en cambio, lo que aumenta de valor es la *información sobre la información*. No es extraño, por tanto, que, invariablemente, en los últimos años, entre las 1(primeras sedes web por número de visitas abunden los directorios como Yahoo (http://www.yahoo.com) y los mo de búsqueda como AltaVista (http://www.av.com) (Alexa, 100Hot).

Desorganización

Los directorios y motores de búsqueda aparecen como los principales agentes de orden en la WWW. Sin embargo directorios y motores de búsqueda cubren una muy pequeña parte: entre un 0,001 y un 16 % y la combinación de l 11 motores de búsqueda no cubre más de un 40 % (Dahn, 2000; Steve y Giles, 1999, y estimaciones propias).

Por si lo anterior fuera poco, hasta un 8 % de los enlaces resultaron erróneos en una investigación a gran escala realizada por el W3 Consortium en 1997. La experiencia nos dice que ese porcentaje, probablemente, no ha hecho que aumentar en los últimos años.

Finalmente, la vida media de un documento en el ciberespacio, y éste es un dato que se mantiene con pocas variaciones en los últimos años, es de 50 días. O sea «hoy está, mañana no está» (W3C y Aguilló, 2000, aunque es último autor relativiza el problema.)

Tampoco parece ser muy fiable la imagen implícita que todos teníamos de una WWW donde casi todo estaba conectado con casi todo. En realidad, un reciente estudio (Broder, 2000) indica que existe alrededor de un 8 % de páginas web totalmente desconectadas. Es decir, hay en la WWW un 8 % de páginas web que ni contienen enlace hacia otras páginas, ni cuentan con ninguna otra página que apunte hacia ellas. Nos podríamos preguntar si respon una estrategia implícita (pero, entonces, ¿para qué publicar en Internet) o a simples errores. Probablemente, obedecerán a una combinación de ambas cosas.

Según ese mismo estudio solamente el 27 % de las páginas de la WWW están intensamente conectadas, formando que los autores denomina el núcleo central de la WWW. En esa zona, dos páginas cualesquiera están unidas a trav de enlaces directos. El mismo estudio revela que existe alrededor de un 42 % de páginas que, o bien apuntan al nú pero no pueden ser accedidas desde el núcleo; o bien son apuntadas por el núcleo pero desde ellas no se puede acc al mismo (el porcentaje, en ambos casos, es similar: un 21 % respectivamente). Así pues, existen amplias zonas aisladas en la WWW si confiamos exclusivamente en la capacidad de los enlaces para movernos por la red.

Pese a todo, existen también algunas notables zonas de orden en la WWW, aunque son de muy pequeño tamaño, c se deben a la actividad de un tipo de organismos, denominados *information gateways,* que detectan, evalúan y describen recursos digitales de interés para la comunidad académica y científica, como, por ejemplo, BUBL (http://www.bubl.ac.uk) y Resource Discovery Network (http://www.rdn.ac.uk).

Los *information gateways* o agencias de evaluación [véase la sede web Desire en http://www.desire.org] son organismos del sector público o privado que realizan labores sistemáticas de descubrimientos de recursos valiosos realizar su correspondiente descripción en sistemas de información de temas genéricos o especializados para pone al alcance de las comunidades a las que sirven; un par de buenos ejemplos de nuestro entorno son Cercador (http://www.cercador.com) e Internet Invisible (http://www.internetinvisible.com).

La moraleja es clara: a todo investigador le saldrá más a cuenta realizar sus búsquedas de información utilizando alguno de tales servicios especializados antes que en directorios o motores de búsqueda para el gran público como Yahoo o AltaVista. Al mismo tiempo, deberíamos promover una mayor actividad en cuanto a la evaluación y descripción de recursos digitales.

Desigualdades

La desigualdad es una de las características más acusadas de Internet, se tome el parámetro que se tome. Empecen por la lengua. El inglés está presente en entre un 80 y un 86 % de los documentos de la WWW. El alemán lo está c 8 %, mientras que lenguas como la francesa, la portuguesa y la española lo están en alrededor de un 3 %. Las dem lenguas del planeta se reparten fracciones ínfimas (Inktomi; OCLC).

Por otro lado, de 6000 millones de personas que pueblan la Tierra, solamente unos 300 millones tienen acceso a Internet y, de ellos, más del 80 % están concentrados en la llamada zona occidental del mundo (Estados Unidos, Canadá y Europa). Los porcentajes son muy parecidos si se examinan los países y las regiones de origen de los servidores web de todo el planeta: Estados Unidos solamente se lleva el 55 % y si se suman a los de Canadá y Eur dan más del 80 % (Global Reach, OCLC).

Finalmente, uno de los datos más espectaculares sobre la desigualdad del ciberespacio la ofrece el último estudio c Alexa (http://www.alexa.com) sobre la WWW, a saber: el 0,5 % de las sedes atrae el 80 % del todo el tráfico de la WWW. En cifras absolutas, de los 5 millones de sedes web de todo el mundo, solamente 15 000 de ellos reúnen el % del tráfico total.

Gratuidad

Las características anteriores no son las únicas, por supuesto. Del ciberespacio o de la WWW se pueden afirmar o muchas cosas que nosotros no hemos podido abordar aquí. Se dice que Internet incrementa las diferencias entre ri pobres en información y, entre ricos y pobres sin más. También se dice que Internet favorece la democracia, la lib y la cooperación.

Pero no queremos dejar de comentar que el ciberespacio ha aportado una más que notable «cultura de la gratuidad que aporta grandes ventajas para los ciudadanos consumidores de información y, es de esperar, para la sociedad en conjunto.

Es cierto que algo gratis no significa algo no financiado: de algún sitio ha de salir la financiación, por ejemplo de l

publicidad; ni necesariamente altruista: los motivos pueden ser perfectamente egoístas, por ejemplo, ganar cuota d mercado.

Además, a veces se minimiza el valor de la información gratuita y se pone en duda su calidad, afirmando que las informaciones realmente valiosas siguen siendo de pago y lo que tenemos en la WWW también se podía obtener d forma gratuita antes por otros medios.

Pero la realidad es que el investigador, el académico o el profesional nunca había tenido, sin coste directo alguno, acceso a tanta información como ahora a través de la WWW. Mencionemos algunos ejemplos notables de fuentes información gratuitas:

• Las dos bases de datos más utilizadas del mundo, Medline en el campo de las ciencias de la vida y Eric en el can de la educación y las ciencias sociales son de consulta gratuita a través de Internet. No es el único caso. Otros ejer notables, del mundo de la comunicación audiovisual son AllMovie e Internet Movie Database.

• Prácticamente la mayoría de los diarios de referencia de todo el mundo pueden consultarse gratis a través de Inte aunque no siempre se puedan consultar números atrasados de más de siete días.

• La Enciclopedia Británica está disponible, también de forma gratuita, en su versión de Internet. A ella se han uni por lo menos otras dos obras de gran prestigio: Encarta y Funk and Wagnalls. En total, son varios cientos los diccionarios y enciclopedias, de todas las ramas del conocimiento, y buena parte de ellas de gran calidad y solven que pueden consultarse de manera gratuita en Internet.

• Una parte importante de las bases de datos de ciencia y tecnología producidas por el Consejo Superior de Investigación Científica (CSIC) sobre la producción científica española puede consultarse de manera gratuita.

Los ejemplos podrían multiplicarse, pero mencionemos solamente dos más de tipo genérico: prácticamente no hay campo del conocimiento que no disponga, por lo menos, de alguna publicación académica digital de calidad de ac gratuito. Finalmente, numerosas instituciones académicas y de investigación, así como los propios autores por su propia cuenta, publican actas de congresos, seminarios, informes, tesis doctorales, etc., de forma gratuita a través d red.

Otros campos también han conocido esta especie de fiebre de la gratuidad: un inmenso número de servidores de Internet está gestionado con Apache, un programa *freeware* (software gratuito), así como existen miles de otras aplicaciones de interés académico y científico, algunas de sorprendente calidad, que son de tipo *freeware* [véase, p ejemplo, Freeware Home en http://www.freewarehome.com, sin mencionar el enorme empuje de Linux, un sistem operativo que empieza a amenazar la cuota de mercado de los grandes, léase Microsoft].

No sabemos cuánto durará esta situación, no sabemos si esa gratuidad ha venido para quedarse o solamente es una forma que tienen las empresas de asegurarse cuota de mercado hoy para tener «clientes cautivos» cuando decidan empezar a aplicar tarifas mañana.

Pero, mientras tanto, lo que resultaría equivocado sería ignorar la riqueza de la oferta de información que podemos encontrar en Internet y que, convenientemente utilizada, puede hacer mucho más fácil la vida de cualquier profesi del mundo de la ciencia y la tecnología.

Conclusiones y prospectiva

La investigación en línea es, al mismo tiempo una actividad y un campo de estudio apasionantes. En el futuro, probablemente, el ciberespacio acaparará la mayor parte de las publicaciones de temas de ciencia y tecnología y su influencia no dejará de crecer.

Se dice que, en los próximos cinco años, la búsqueda de información en la WWW estará resuelta gracias a los nue

motores de búsqueda y al uso de herramientas de procesamiento del lenguaje natural. Ojalá sea cierto, pero, en toc caso, nos conviene ahora y nos convendrá aún más en el futuro conocer cómo es el ciberespacio en realidad, cuál forma y el mejor modo de explotar sus características para conseguir tener mayor calidad de vida e intentar hacer este planeta un lugar mejor y más equilibrado donde vivir (¿para qué diablos sino querríamos el ciberespacio?).

Por último, hay una prospectiva que no me resisto a citar: según R. Kurzweill, en el 2019, un ordenador de mil dó (o sea, de unas 180 000 pesetas) tendrá el poder de procesamiento de un cerebro humano, y hacia el 2029, un típic microordenador tendrá el poder de procesamiento de mil cerebros (R. Kurzweil, 1999, citado por Sherman, 2000).

De entrada puede asustar un poco, porque todos recordamos al ordenador HAL de *2001, Odisea del espacio* pero piensa bien quizás no sea una perspectiva tan mala: quizás así, finalmente, habrá vida inteligente en este planeta...

Fuentes de información

Bibliografía

Ackermann, E.; Hartman, K. *The information specialist's guide to searching and researching on the Internet and World Wide Web,* Wilsonville, ABF Content, 1998.

Aguillo, I.F.: «Herramientas de segunda generación». En: Cid, P.; Baró, J. (eds.): *Anuari SOCADI de Documenta Informació*. Barcelona, SOCADI, 1998: 85-112.

Aguillo, I.F.: «Contenidos de I+D en Internet: mitos y leyendas», *Mundo Científico*, abril 2000: 22-25.

Basch, R.: *Researching online for dummies*, Foster City, IDG Books, 1998.

Broder, A. *et al.*: *Graph structure in the web*, http://www9.org/w9cdrom/160/160.html, 2000.

Burdoncle, F.; Bertín, P.: «Buscar agujas en un pajar», *Mundo Científico*, abril 2000: 58-63.

Codina, Ll.: «El ecosistema informativo de la WWW», *Datamation* 2000; 166 (mayo): 42-46.

Codina, Ll.: «Evaluación de recursos digitales en línea: conceptos, indicadores y métodos», *Revista Española de Documentación Científica* 2000; 23 (1): 9-44.

Dahn, M.: «Counting angels on a pinhead: critically interpreting web sizes estimates», *Online* 2000; enero, y http://onlineinc.com/onlinemag/OL2000/dahn1.html.

Glossbrenner, A.; Glossbrenner, E.: *Search engines for the world wide web,* Berkeley, Peachpit Press, 1998.

Hock, R.: *The extreme searcher's guide to web search engines: a handbook for the serious searcher,* Medford, CiberAge Books, 1999.

Lynch, C.: «La exploración de Internet», *Investigación y Ciencia* 1997; mayo: 38-43.

Miller, P.; Greenstein, D. (eds.): *Discovering online resources across the humanities: a practical implementation the Dubin Core,* Londres, University of Bath, The UK Office for Library and Information Networking, 1997.

Sherman, Ch.: «The future of web search», *Online* 1999; 23 (3): 54-61.

Sherman, Ch.: «The future revisited: what's new with web search», *Online* 2000, http://www.onlineinc.com/onlinemag/OL2000/sherman5.html

Sulivan, D.: «Crawling under the hood: an update on search engine techonology», *Online* 1999; 23 (3): 30-36.

Sitios web

Alexa: http://www.alexa.com

Desire: http://www.desire.org

Dublin Core: http://purl.org/dc

Inktomi Web Map: http://www.inktomi.com

Internet Invisible: http://www.internetinvisible.com

Invisible Web: http://www.invisibleweb.com

Global Reach: http://www.glreach.com

OCLC: http://www.oclc.org

OMNI: http://www.omni.ac.uk

Ressource Discovery Network: http://www.rdn.ac.uk

Search Engine Watch: http://www.searchenginewatch.com

Search IQ: http://www.searchiq.com

W3C: http://www.w3c.org

Lluís Codina

Doctor en ciencias de la información y profesor de documentación en los medios y documentación periodística. Responsable de la sección científica de Biblioteconomía y Documentación, del Departamento de Ciencias Política Sociales, de la Universitat Pompeu Fabra. Es miembro del Observatorio de la Comunicación Científica y del Insti de Lingüística Aplicada y vicepresidente del Colegio Oficial de Bibliotecarios-Documentalistas de Cataluña. Es codirector del proyecto Documentación ~Digital de formación a distancia utilizando hipertextos y tecnología web (http://docdigital.upf.es). Ha publicado más de doscientos artículos sobre temas de su especialidad y es autor de *E libro digital* (Barcelona, 1996), cuya nueva versión está prevista para principios del 2001.

lluis.codina@cpis.upf.es

[Foto carnet: sí]

NOTA **Capturas de pantalla PARA ADORNAR EL TEXTO y sus pies correspondientes**

http://www.oclc.org/oclc/research/projects/webstats/index.htm

El consorcio OCLC mantiene un proyecto de caracterización de la WWW y publica los informes en su página wel

http://www.internetinvisible.com

Gracias a recursos como Internet Invisible, podemos conocer cuáles son las mejores bases de datos españolas de ciencia y tecnología disponibles a través de la WWW

http://www.rdn.ac.uk

Ressource Discovery Network proporciona información sobre bases de datos y otros recursos de información en ciencia, en ciencias sociales y en humanidades

http://www.desire.org

Desire es un proyecto de la UE sobre temas de investigación en línea y evaluación de recursos digitales. Publica informes y estudios de gran valor sobre el campo