Bibliotecas digitales: retos y libre acceso

María Mercedes Martínez González
Universidad de Valladolid

0.1. Resumen

Se estudia el impacto en las bibliotecas del uso de soportes electrónicos, procesos automáticos y de la Web para difundir información. Se analiza el acceso a la información digital, atendiendo especialmente a los modelos económicos de explotación, así como los problemas relacionados con la selección, la evaluación de la calidad y la preservación de dichainformación.

Palabras clave: Bibliotecas digitales. Modelos económicos. Libre acceso. Búsqueda de información. Calidad. Preservación.

0.2. Abstract

How libraries are influenced by the use of electronic resources, automatic processes and the Web to publish information is examined. Access to digital information, with a focus in the exploitation economic models, selection, information assessment and preservation issues are considered.

Keywords: Digital libraries. Economic models. Open access. Information seeking. Quality. Preservation.

1. Introducción

¿Qué hace útil a una biblioteca digital? ¿Qué aportan estos sistemas a las bibliotecas tradicionales? El primer argumento a favor de su utilidad es la capacidad de los sistemas informáticos para albergar grandes cantidades de información a *bajo coste*. El precio de los soportes informáticos (discos duros, CD-ROM) es considerablemente más reducido que el precio de una biblioteca tradicional. Debe tenerse en cuenta que el coste del edificio que alberga la biblioteca suele ser elevado. El desgaste físico de los ejemplares que contiene implica un coste adicional en renovación que con las versiones electrónicas puede minimizarse. Otro argumento quizás aún más poderoso a favor de estos sistemas es la posibilidad de ofrecer un *acceso masivo* a la información. La disponibilidad de un ejemplar en una biblioteca tradicional está condicionada al hecho de que no haya otro usuario que lo tenga en préstamo. El acceso a los ejemplares es "uno a uno", secuencial: dos usuarios no pueden acceder

simultáneamente al mismo ejemplar; y, para que uno de ellos pueda disponer de él, debe esperar a que sea devuelto. La capacidad de los sistemas informáticos de ofrecer información mediante sistemas tan accesibles y populares como la Web facilita que varios usuarios puedan acceder simultáneamente a la misma información, sin necesidad de esperar a que otros usuarios "liberen" dicha información.

De hecho, éste es el primer aspecto que recibe atención en este trabajo, puesto que quizás es el mayor reto al que se enfrentan las bibliotecas: continuar la labor que han desempeñado siempre como difusoras y proveedoras de información, facilitando el acceso a ésta al mayor número posible de individuos. La tecnología aporta los medios para proporcionar el más amplio y fácil acceso que se haya conocido nunca. Sin embargo, su impacto en la economía (términos del tipo *e-commerce*, *e-business*, están en pleno auge) hace que todo lo digital (todo aquello a lo que se le pueda anteponer una "e-") sea objeto de fuertes intentos de establecer medidas "protectoras" de ciertos derechos como los de propiedad intelectual. Ocurre que estas medidas resultan en muchos casos más restrictivas aún —aquí surge el punto conflictivo— que las que se han aplicado a los soportes tradicionales. En el apartado tercero se reseñan algunos de los modelos económicos propuestos para las bibliotecas digitales. Algunos de ellos se inspiran en modelos ya conocidos y ampliamente probados en las bibliotecas. El libre acceso, por ser el que parece preferir la comunidad académica y científica, y el que más se puede beneficiar de los avances tecnológicos, es el que recibe más atención en este trabajo.

Por otro lado, las bibliotecas digitales permiten automatizar algunos procesos propios de una biblioteca. Pero, ¿son capaces las herramientas informáticas de realizar estas tareas de modo autónomo, sin que sea necesaria una revisión posterior del experto? ¿Lo serán en el futuro? El uso creciente de soportes electrónicos y de la Web para difundir información introducen variantes que requieren una revisión de los modos tradicionales de resolver ciertas tareas. En la sección cuarta se revisan tres de estas funciones. La primera es la búsqueda de información. Garantizar la calidad de la información disponible es la segunda. Por último, se hablará de la preservación de la información. Los ítems digitales reciben especial atención, ya que la cantidad de información disponible *únicamente* en formato electrónico es cada vez mayor.

2. Las bibliotecas y la Web

Las bibliotecas digitales ofrecen acceso a sus usuarios a colecciones organizadas de información, siendo ésta su principal diferencia con otros sistemas como la Web. No obstante, ésta última es utilizada por muchos usuarios como una gran biblioteca. Es innegable que la Web se ha convertido en la forma de obtener información preferida por más usuarios. No toda la información disponible tiene

valor, e incluso se pueden encontrar enormes cantidades de información sin ningún interés. Sin embargo, entre la información disponible es posible encontrar alguna información de buena calidad; por ejemplo, una proporción importante de artículos científicos están disponibles, bien a través de la página Web de su autor, de la institución a la que pertenecen, etc. Y según parece, cada año que pasa, el porcentaje de información interesante aumenta. El resultado es que los usuarios muestran una cierta tendencia a utilizar la Web, un sistema de libre acceso, como una importante fuente de información o referencias. En consecuencia, la información disponible en este medio (ítems digitales) debe ser considerada tanto al plantearse dónde encontrar información (información en la Web), como al hablar de calidad (no hay revisiones que garanticen la calidad de la información) o de preservación (información disponible únicamente en la Web).

3. Modelos económicos para el acceso a la información

¿Debe estar abierto el acceso a la información al mayor número posible de usuarios, o debe, por el contrario, restringirse? Con los soportes electrónicos y la difusión masiva de la Web surgen nuevos elementos que se deben considerar a la hora de responder a esta pregunta. A continuación se revisan los tres modelos económicos más importantes propuestos para las bibliotecas digitales, y se termina con una breve reflexión sobre cómo influyen algunos aspectos legales en la elección de un modelo u otro.

3.1. Libre acceso

3.1.1. ¿Quién está interesado en proporcionar libre acceso?

Una de las presunciones que se pueden hacer erróneamente es que los proveedores de información siempre están interesados en un acceso restringido a ella. En algunos casos no es así: El proveedor está interesado en proporcionar libre acceso porque esto le permite *promocionarse*. Es el caso de las versiones beta de algunos productos software. También es el modelo que han utilizado algunos portales en Internet para financiarse. En otras ocasiones, la información es *del Estado*, y, por tanto, es pública. Este es el caso de la legislación nacional —aunque el estado español no proporciona acceso gratuito a todas las publicaciones oficiales— o también el de la legislación de la Unión Europea. Finalmente, la *investigación académica* es un entorno en el que los autores suelen estar interesados en proporcionar acceso abierto a sus trabajos, ya que esto aumenta sus posibilidades de ser referenciados por otros autores. El mayor interés en una publicación no es el beneficio económico que de ella se pueda obtener, sino las referencias, que dan prestigio al autor dentro de la comunidad académica.

3.1.2. ¿Cómo se financia el libre acceso?

Alguien debe financiar los servicios que se ponen a disposición de los usuarios. Incluso la más barata de las opciones tiene un coste en material, personal y mantenimiento. Luego, la pregunta lógica es: ¿quién asume este gasto cuando el acceso es libre? A continuación se presentan los modos de financiar el libre acceso más habituales. Se comentan además para cada uno de ellos algunas cuestiones sobre sus posibilidades de desarrollo futuro.

1. *Publicidad*. Es la opción escogida por algunos portales temáticos y buscadores de Internet (*Web search engines*). Esta solución es relevante, porque estas herramientas se han convertido en una importante fuente de información para los estudiantes (los consumidores futuros de información), más cuanto que les permite obtener información actualizada con poco esfuerzo. Pongamos como ejemplo un caso real: Una estudiante de 14 años tenía que buscar información sobre los problemas de integración de la comunidad musulmana en su ciudad para un trabajo escolar. Esta estudiante empleó como principal fuente de documentación Internet, y se apoyó en algunos de estos motores de búsqueda para encontrar documentación. También utilizó los recursos que la biblioteca pública le ofrecía, de la cual tomó prestados una serie de libros. Sin embargo, la información más reciente la consiguió en Internet. Obviamente, esta documentación, que ilustra la situación actual en el momento en que realizaba el trabajo, era especialmente valiosa para ella. ¿Seguirán utilizando este tipo de financiación los portales y buscadores? ¿Deberán combinarlos con otros modos de financiación? ¿O finalmente el modelo será desechado, ya que según parece esta opción no ha dado los resultados que inicialmente se esperaban en Internet?

2. *Inscripciones*. Algunas publicaciones están financiadas por las inscripciones realizadas por los socios de la organización que las mantienen. Un ejemplo es la serie RFC (*Request For Comments*) de Internet (1). En estos documentos se fijan directrices sobre protocolos de red, programas y otros conceptos aplicables a Internet. Los autores envían sus propuestas y el editor asume la tarea de revisión, decidiendo cuáles se publican y cuáles son los cambios que deben hacerse antes de su inclusión en la serie. El editor es miembro de la Internet Society (ISOC), una agrupación profesional, que asume los gastos de mantemiento y gestión de la serie RFC. Esta solución parece adecuada para las conferencias académicas, que podrían utilizar los costes de inscripción para financiar el acceso abierto a sus actas. De hecho, algunas conferencias (2) ya proporcionan este acceso de modo gratuito en Internet.

3. *Financiación pública y donaciones*. El estado puede ser una importante fuente de financiación. Algunos proyectos relevantes están financiados con dinero público, por ejemplo el proyecto Genoma (3). Por otro lado, gran parte de la

investigación se financia de este modo, y, en consencuencia, algunos proyectos publican sus resultados en servidores Web.

4. *La organización productora es financiadora*. En este caso es la propia organización que produce la información la que asume los costes de su publicación. Este es un caso muy común en la Web. Por ejemplo, es lo que ocurre con los estándares que publica el World Wide Web Consortium (W3C). Esta organización se encarga de elaborar las normas destinadas a regular distintos aspectos relacionados con la Web (arquitecturas de software, modelado de información, protocolos, etc.). Su publicación de acceso gratuito —para facilitar así su difusión— se realiza a través del propio servidor Web del W3C (4), de cuyo mantenimiento se encargan las entidades que componen este consorcio. Pero este modelo no tiene por qué ser exclusivo del entorno tecnológico. ¿Por qué no podrían asumir una responsabilidad similar las universidades, que son, a fin de cuentas, productoras de información valiosa a través del trabajo de sus miembros (tesis doctorales, resultados de investigación, trabajos docentes y otros). Esto no tiene por qué ser una posibilidad del futuro; iniciativas como el portal Universia, que aglutina sesenta y cuatro universidades españolas e hispanoamericanas (5), pueden suponer el primer paso en este sentido.

5. *Los autores son editores*. Como se comentó anteriormente, los investigadores suelen estar muy interesados en facilitar el acceso a sus trabajos. Los grupos de investigación suelen asumir la responsabilidad de publicar en sus servidores las publicaciones de sus miembros. Incluso en algunos casos, cuando las publicaciones finalizan en revistas con normas restrictivas respecto al acceso, estos grupos de investigación publican los borradores que los investigadores utilizaron antes de redactar los artículos definitivos. Lo esencial de su trabajo está disponible en dichos borradores, con lo cual se aseguran la posibilidad de ser referenciados por otros investigadores (figura 1). No obstante, debe tenerse en cuenta que la publicación científica no tiene coste cero.

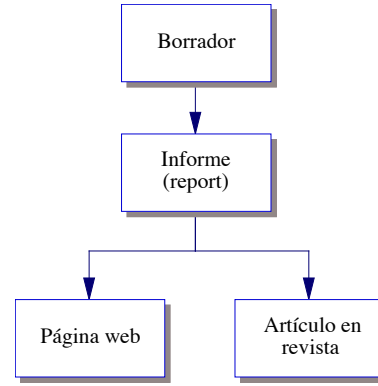

Fig. 1. La publicación científica de los resultados de un trabajo de investigación bifurca en ocasiones la redacción inicial de un borrador en varios documentos para garantizar simultáneamente la máxima difusión (página Web) y la máxima calidad (publicación en revista).

3.2. Préstamo interbibliotecario (ILL)

El préstamo interbibliotecario (*InterLibrary Loan, ILL*) es un modelo bien conocido en las bibliotecas. Permite

compartir libros, microfilms, copias de artículos en revistas, capítulos de libros, etc. El préstamo, como su propio nombre indica, se realiza entre bibliotecas. Es la biblioteca la que actúa como intermediario a favor del usuario que solicita un documento. Al tratarse de un modelo con una larga tradición tiene la ventaja de estar regulado legalmente. En cuanto al modo de financiar este servicio, se distinguen básicamente dos posibilidades:

- *La biblioteca financia el coste*: En estos casos el préstamo es considerado un servicio básico que se proporciona a los usuarios. Suele ocurrir también que el coste de estos préstamos es un porcentaje pequeño respecto a los demás gastos de la biblioteca.
- *El usuario paga el coste del préstamo*: Cuando recibe el documento requerido, debe abonar a la biblioteca el precio que ésta pagó a su vez por él.

Dado que éste es un modelo bien conocido, es lógico preguntarse si es adaptable al caso de las bibliotecas digitales. Y, en caso de que sea así, ¿requiere alguna consideración especial el hecho de trabajar con ítems digitales? La respuesta es que sí es posible aprovechar este modelo con los documentos electrónicos, y de hecho se está haciendo. Sin embargo, existe una dificultad: el coste de plantilla. El personal que se ocupa de encontrar el documento que el usuario solicita realiza búsquedas en los catálogos para localizar la biblioteca que dispone del ítem requerido. Esto incluye los catálogos basados en la Web, y aquí surge el problema: el esfuerzo necesario para localizar un ítem en este tipo de medio —donde hoy en día los catálogos distan de estar bien organizados, de ser completos, y donde además cada vez más recursos únicamente están disponibles en la Web— no parece sostenible. Será necesario, pues, buscar soluciones a este problema.

3.3. Préstamos a los usuarios

En este caso, a diferencia del préstamo interbibliotecario, se asume que la biblioteca local posee o tiene las licencias oportunas sobre el material solicitado por el usuario. Se trata entonces de adaptar el modelo de préstamo temporal al caso de los ítems digitales: el documento que la biblioteca proporciona al usuario debe tener una disponibilidad limitada en el tiempo, de igual modo que el préstamo de un libro ha supuesto siempre que, transcurrido un cierto período, el usuario debía retornarlo a la biblioteca. La tecnología imita aquí el funcionamiento tradicional. El documento electrónico que el usuario descarga en su disco duro tiene asignado un período de validez, transcurrido el cual dicha copia deja de ser accesible. Existen varios productos software que implementan este tipo de modelo, asociado al concepto de libro electrónico. Algunos ejemplos son netLibrary (6), eBook de Adobe Inc. (7), o el Lending Library Format (8). Este modelo no plantea más particularidades en su adaptación que las meramente tecnológicas; y éstas, como deja claro la disponibilidad de software que lo implementa, están resueltas.

3.4. Aspectos legales

A la hora de decidir qué modelo adoptar se deben tener en cuenta los aspectos legales, relacionados con la propiedad intelectual, que restringen en algunos casos la posibilidad de ofrecer un acceso completamente libre a los documentos. En este apartado no se tratan aspectos generales —que, por otro lado, quedan fuera del alcance de este trabajo—, sino que se considera brevemente aquéllo que es particular al caso de los documentos electrónicos y que puede afectar a las decisiones de diseño e implementación en una biblioteca.

Hay disconformidad sobre si los documentos electrónicos son equiparables a otros tipos de documentos y, por tanto, si se deben permitir los mismos derechos de acceso. De hecho, parece que algunos editores los consideran distintos y en consecuencia construyen barreras (legales y tecnológicas) que dificultan a los usuarios usar estos materiales con la naturalidad con que están acostumbrados a hacerlo con otros tipos de documentos. Normalmente, una vez ha pagado por el original, un usuario puede hacer copias del material impreso para su uso particular. Un caso bien conocido es el del formato MP3 (9) (Arms 1999), que, si bien en sí mismo no es ilegal, facilita la distribución de copias piratas de música, ya que las copias no tienen ninguna pérdida de calidad respecto al original (10). Sin embargo, la tecnología desarrollada para dar soporte a algunos formatos electrónicos tiene como objetivo dificultar la realización de copia alguna de dicho documento, sea cual sea el objetivo de quien pretende copiarlo. Esto también repercute en el entorno educativo, donde tradicionalmente se han utilizado copias de documentos para facilitar la tarea docente, y donde este tipo de tecnología impediría este método de trabajo. Esta cuestión puede afectar además a las iniciativas tendentes a garantizar la preservación de la información que se encuentra disponible en formato electrónico, como el proyecto *Internet Archive* (Kahle, Prelinger y Jackson, 2001), cuya portada de Internet se presenta en la figura 2. ¿Tienen o no estos proyectos derecho a almacenar copias de los documentos disponibles en Internet en sus colecciones? Lo mismo se puede decir de los buscadores de Internet (Google, AlltheWeb...). Si bien quedan aspectos por regular, ya existen normativas que abordan específicamente el tema de los derechos de propiedad intelectual cuando se trata de procesos tecnológicos (11), y que aportan una cierta esperanza a aquellos que esperan que la Web sea, entre otras cosas, la mayor y más accesible fuente de información disponible. Herramientas como Google, usada por tantos usuarios para localizar información, tienen garantizada de este modo la legalidad de sus servicios.

4. Retos de las bibliotecas digitales

En las bibliotecas tradicionales el coste se reparte entre los documentos; los edificios, mobiliario, etc.; y el personal. En las bibliotecas digitales los costes son, por su parte, documentos, ordenadores y redes informáticas, y personal. Un

Fig. 2. Internet Archive construye una biblioteca digital a partir de los sitios Web. Al igual que las bibliotecas tradicionales, proporciona libre acceso (gratuito) a los investigadores, historiadores, estudiantes y público en general. Actualmente más de diez billones de páginas Web están disponibles en la biblioteca, accesible en http://www.archive.org/.

porcentaje importante de los gastos de las bibliotecas se dedica a personal. Pero si en el futuro han de ampliar considerablemente la cantidad de información que ofrecen, no parece razonable esperar que los costes de personal aumenten de modo directamente proporcional a la información disponible. Dicho de otro modos, estos costes tendrán que "reducirse". ¿Pueden las bibliotecas digitales ayudar a optimizar este gasto?

Si bien los procesos automáticos no son especialmente elaborados o "inteligentes", un factor juega a su favor: la capacidad de los ordenadores de ejecutar miles de operaciones en tiempos mínimos. Esta capacidad puede hacer que algoritmos simples (los conocidos como métodos de fuerza bruta) den resultados satisfactorios allí donde se podría suponer que no sería así. Un ejemplo es Google, que en marzo de 2000 (12) soportaba 8,5 millones de búsquedas diarias con una plantilla de 85 personas, de las cuales 14 eran doctores y la mitad técnicos, y con una infraestructura compuesta de 2.500 PCs con Linux y 80 terabytes de disco. Estos datos son elocuentes: el número de búsquedas efectuadas es impresionante, si se tiene en cuenta lo reducido de la plantilla. La capacidad de cálculo justifica por si misma el interés de los sistemas automáticos.

Pero una vez que esto no se pone en duda, hay que preguntarse ¿qué más son capaces de aportar estos sistemas? ¿dónde están sus limitaciones? ¿dónde se debe investigar para mejorarlos? En definitiva, ¿cuáles son sus retos más importantes? El reto es "imitar" al experto, simular su experiencia. Los sistemas que lo hagan deben automatizar algunas de sus tareas, obteniendo resultados lo más parecidos posible a aquéllos que el experto habría obtenido, para que se consideren útiles. El personal de una biblioteca aprovecha su experiencia y cualificación para seleccionar material (documentos), catalogar e indexar estos materiales, buscar información, evaluarla para determinar su grado de calidad, y actuar como un servicio de referencia para los usuarios, ayudándoles a localizar la información que buscan. Actualmente se automatizan algunos de estos procesos, aunque sea parcialmente, como se puede apreciar en la indexación automática de los documentos que realizan Lycos, Infoseek, Altavista, Google y los buscadores de Internet en general; la ordenación de los documentos de la respuesta a una consulta en función de su calidad que efectúa Google —cuyo algoritmo se comenta más adelante—; o la creación automática de colecciones que implementa CiteSeer (Lawrence, Giles y Bollacker 1999).

En general, se puede diferenciar tres tipos de funciones susceptibles de automatización que conviene evaluar en cuanto al grado de automatización conseguido en la actualidad y al impacto que los avances tecnológicos y la utilización de formatos electrónicos y estrategias de difusión basadas en la Web pueden tener en ellos. Dichos procesos básicos son:

1. *Encontrar información*. Los usuarios buscan información. Sus búsquedas son más fáciles cuando se utilizan catálogos e índices. Tradicionalmente se usan los catálogos creados por expertos en la materia, con un grado de habilidad alto, lo cual los encarece. Esto impide utilizar masivamente este tipo de catálogos. Alternativas con un grado de automatización mayor son más baratas y, por tanto, las elegidas en los casos donde no se puede asumir el coste de los expertos.

2. *Calidad* de la información. Tradicionalmente los editores y las bibliotecas eran los responsables de evaluar y garantizar la calidad. En Internet, donde una cantidad importante de documentos los publican directamente sus autores, es difícil distinguir las fuentes de información adecuadas de las que no lo son; ya no es posible contar con la revisión de expertos en la materia. Los métodos de evaluación tradicionales —basados en la confianza en los procesos de revisión— dejan de ser la garantía de calidad.

3. *Perdurabilidad*. Mucha información en la Web puede estar disponible en servidores temporales, de modo que posteriormente puede resultar inaccesible. En ocasiones, cada vez con más frecuencia, ésta es además la única forma en que

se publica esa información. Si esta información es importante, sería conveniente preservarla. ¿Cómo conseguirlo?

4.1. Encontrar información

Encontrar los documentos que contienen la mejor información no es una tarea fácil. En este apartado se comentan algunas de las estrategias utilizadas en la automatización de estas tareas, y cómo se mide la calidad de estos procesos automáticos. La automatización más "tradicional" de esta tarea es la creación de índices en línea como los conocidos OPACs, o servicios como Medline, Science Citation Index. La característica común a todos ellos es que los índices y catálogos han sido construidos cuidadosamente por expertos. Por otro lado, también ocurre que en muchos casos aquéllos que realizan las búsquedas son usuarios expertos, que saben formular las consultas del modo más adecuado para obtener los mejores resultados. Su desventaja es que el alto grado de especialización del personal que los crea y los utiliza los hace costosos. Otras estrategias tratan de suplir o abaratar estos costes de varias formas:

- Mediante la *creación automática de índices*. La indexación automática es uno de los procesos automáticos con más tradición, si bien se siguen proponiendo nuevas aproximaciones, tendentes a mejorar la eficacia de estos índices y adaptarlos a los nuevos formatos disponibles (documentos XML, imágenes, vídeo...).
- Calculando índices de *similaridad entre consultas* para así ajustar las consultas de los usuarios noveles a consultas más efectivas de usuarios expertos. Si la consulta que realiza un usario novel es similar a alguna consulta realizada previamente por un experto en el tema objeto de la búsqueda, se utiliza la consulta del experto, que se presupone más precisa y, por tanto, se espera proporcione mejores resultados.
- *Refinando las búsquedas* aprovechando los resultados de búsquedas anteriores. Estas propuestas consisten esencialmente en identificar los atributos más relevantes de aquellos ítems de la respuesta inicial que han interesado al usuario, y redefinir la consulta utilizando esos atributos. La base de estas propuestas es que son los atributos de los objetos mejor calificados por el usuario los que éste usaría si tuviese que redefinir la consulta. Esta aproximación se ha usado sobre todo en la búsqueda de imágenes, donde se solicita al usuario que seleccione aquéllas que más le interesan, en vez de pedirle que formule directamente una consulta por atributos.

4.1.1. Calidad de los procesos automáticos

Normalmente se evalúa la bondad de los procesos automáticos mediante cálculos estadísticos sobre las respuestas obtenidas a las consultas. Es decir, se mide

la calidad de la respuesta para juzgar la calidad del proceso que la proporciona. Los dos índices más usuales son la *exhaustividad (recall)* —porcentaje de documentos relevantes recuperados respecto al total de documentos recuperados— y la *precisión (precision)* —porcentaje de documentos relevantes recuperados respecto al total de documentos relevantes disponibles. El Anexo 1 muestra un ejemplo de cálculo de estos índices.

Cuando no se controla las colecciones —como en la Web, donde no se sabe cuántos documentos hay disponibles y, por tanto, no se puede saber cuántos son relevante— únicamente la exhaustividad es aplicable. La precisión no se puede calcular. Esto lleva a la búsqueda de nuevos criterios para estimar la calidad de una respuesta, basados más en los propios usuarios que en la mera utilización de técnicas estadísticas. Estos criterios están directamente ligados alas estrategias de mejora de consultas comentadas en esta sección, en las que se prioriza la información (conocimiento) que el usuario proporciona al sistema sobre las respuestas obtenidas; se trata de la aplicación de técnicas de *soft computing* (*data mining*, *fuzzy logic*) a los procesos de recuperación de información.

4.2. Calidad

El usuario busca trabajos de calidad, que pueda considerar buenas referencias. Tradicionalmente los editores y las bibliotecas asumían esta tarea. Sin embargo, en Internet la situación cambia: son los propios autores quienes dejan disponibles los documentos que crean. No hay ninguna revisión que garantice al usuario la calidad de estos ítems. ¿Es posible encontrar soluciones que ayuden al usuario a distinguir entre aquéllo que es adecuado y lo que no lo es? A continuación se revisan los criterios que usan los usuarios para reconocer la calidad de un trabajo, y cómo se aplican (o no) estos criterios en Internet.

1. Se mantiene un criterio tradicional: el *editor* es la garantía. Además, el usuario utiliza la *apariencia* del trabajo o publicación como pista indicativa de su calidad. Las figuras 3 y 4, en la página siguiente, muestran un ejemplo de publicaciones electrónicas donde este criterio falla. El *Journal of the ACM* (13) es una revista prestigiosa en el entorno informático, una garantía de calidad. Por otro lado, la conferencia sobre *Digital Libraries* de la ACM no tiene el mismo estricto proceso de revisión de trabajos, con lo cual la garantía que aporta no es comparable. Sin embargo, la apariencia externa es la misma.

2. El editor selecciona los autores. La reputación del editor determina la calidad. Si no hay un editor cuyo prestigio sirva de garantía, situación común en la Web, este criterio no es aplicable.

3. *El editor selecciona revisores externos*. Una vez más, el lector basa su evaluación en la confianza que tiene en el editor. La observación en este caso es la misma que en el anterior.

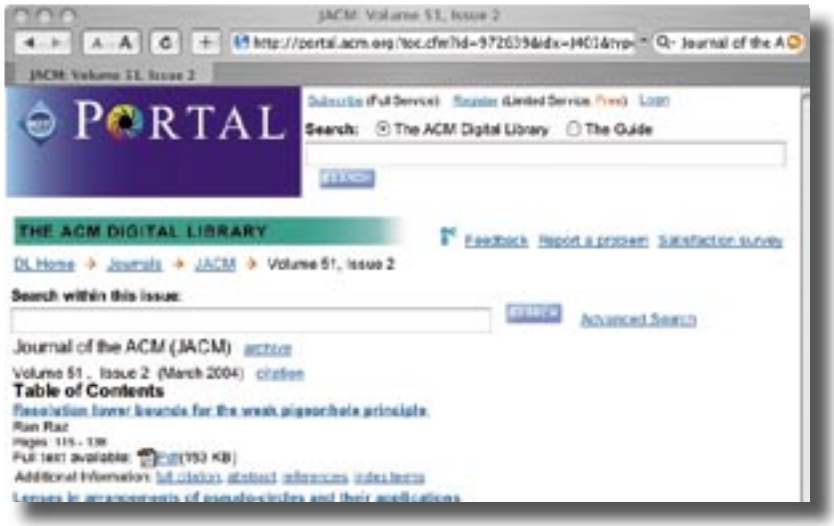

Fig. 3. Journal of the ACM.

Fig. 4. International Conference on Digital Libraries de la ACM.

4. *Revisores independientes*. La valía de la revisión depende de la reputación de la revista o medio en que se publica y de lo bien que se haga la revisión. La tecnología digital facilita la puesta en funcionamiento de este tipo de estrategia:

quienes desean evaluar un ítem sólo tienen que hacerlo a través de una página Web, con unos cuantos clicks de ratón (la figura 5 muestra la herramienta de evaluación utilizada en Amazon). La ventaja de este método en Internet es que se pueden tener muchas revisiones con poco esfuerzo por parte de los usuarios. No obstante, hay que decir que los usuarios son "vagos": en muchos casos prefieren ahorrarse el tiempo de una evaluación, por pequeño que sea. Prueba de ello son las pocas evaluaciones (ninguna en ocasiones) que tienen algunos libros que Amazon vende —y permite evaluar— a través de su portal.

Fig. 5. Evaluación mediante revisores independientes en amazon.com (cuadrante inferior izquierdo).

5. *Métodos estadísticos* de evaluación de la calidad. La calidad se computa automáticamente. Así, por ejemplo, mediante el algoritmo PageRank Google calcula la calidad de una página Web en función del número de enlaces que apuntan hacia ella, ponderando automáticamente la calidad de cada enlace (Brin y Page 1999). Obviamente, este criterio es discutible, pero su gran ventaja es que es completamente automático.

Sobre calidad y los procesos de revisión se pueden considerar otros aspectos, como el modo en que el propio proceso de revisión se ha visto afectado por el uso de los soportes electrónicos para publicar revistas y documentos en general. El lector interesado en este tema puede encontrar información en (Weller, 2001).

4.3. Perdurabilidad

Cada vez más información relevante está disponible en la Web; y, en algunos casos *únicamente* en la Web. Ocurre además que las organizaciones que mantienen dicha información pueden desaparecer con el tiempo, con lo cual las páginas

que contienen esa información también desaparecen. Por lo tanto, el reto que se plantea es cómo garantizar la perdurabilidad de dicha información, para que pueda ser accesible también en el futuro. Ya existen algunas iniciativas cuyo objetivo es éste. La figura 1 muestra el portal de *Internet Archive*, que está construyendo una biblioteca digital a partir de los documentos encontrados en los sitios Web. La estrategia seguida en este tipo de iniciativas se basa en aprovechar la potencia de cálculo y almacenamiento de los ordenadores: periódicamente se toman "snapshots" de la información disponible, copiándola tal cual se encuentra disponible en ese momento. La estrategia es sencilla. No obstante, existen algunas decisiones importantes que se deben tomar, que constituyen el auténtico desafío, y que se enumeran a continuación:

1. *Qué sitios almacenar*. Previamente se comentó que no toda la información disponible en Internet es de buena calidad. Tampoco toda interesa, independientemente de su calidad. Según la intención de quienes realizan las tareas de preservación, habrá temas relevantes y temas sin interés. Por ejemplo, si se trata de preservar documentos relacionados con la pintura, se descartarán trabajos sobre informática, por muy importantes que éstos sean. Se distinguen dos modos de actuar: preservar todo lo que se encuentra dentro de una cierta categoría, o sólo aquellos sitios que ha seleccionado un experto. La diferencia es evidente: en el primer caso se optimiza el coste del proceso de preservación si bien no se garantiza la calidad de lo preservado; mientras que en el segundo se garantiza la calidad, aunque los costes iniciales se encarecen por la intervención de los expertos.

2. *Frecuencia de los "snapshots"*. Las copias pueden hacerse con una frecuencia semanal, mensual, etc., en función de las circunstancias. Los cambios entre una copia y la siguiente dependerán de dicha periodicidad. Los costes de almacenamiento serán mayores a mayor frecuencia, aunque el seguimiento de la evolución de la información será, lógicamente, mejor.

3. *Qué almacenar*. Se puede optar por copiar documentos completos (páginas HTML, por ejemplo), texto e imágenes únicamente, sólo metadatos (por ejemplo, metatags HTML), etc. Qué es realmente importante, digno de perdurar, es algo que debe decidirse en cada caso particular. La tecnología aporta los medios para implementar cualquiera de estas opciones, pero no los criterios para tomar una decisión.

4. *Aspectos legales*. Anteriormente, se comentó la influencia de los aspectos legales sobre los modelos económicos. Estos aspectos también afectan a la perdurabilidad. Al copiar información disponible en la Web, se debe estar seguro de que hacerlo sea legal. La falta de regulación específica sobre este tema plantea serias dudas a quienes afrontan este reto. Mientras no exista regulación más específica, se presume que hay permiso para copiar cualquier página Web, salvo que expresamente se indique lo contrario. Actualmente las herramientas que recorren

automáticamente los servidores Web y copian sus páginas —robots (14)— reconocen una negación de dicho permiso si encuentran en un servidor un fichero robots.txt. Este fichero que se ajusta a un estándar, cuya especificación está accesible en http://www.robotstxt.org/, contiene un listado de páginas sobre las que se deniega el permiso de copia. Los robots reconocen y procesan automáticamente la información de este fichero.

5. *Identificación*. Debe decidirse también cuál es el mejor modo de identificar los documentos almacenados. La URL en la que se ha encontrado es una opción, pero puede ser poco significativa en cuanto que aporta poca información sobre el documento. Otras opciones tienen en cuenta aspectos más descriptivos, como la fecha en que el documento fue creado o copiado, o la secuencia histórica en los cambios hechos al documento.

6. Cómo *garantizar la accesibilidad futura*. Los formatos utilizados actualmente pueden quedar obsoletos a medida que se produzcan avances tecnológicos. ¿Se deberá entonces migrar los ficheros almacenados a los nuevos formatos? ¿Será preferible (y posible) hacerlo de modo totalmente automático? ¿O será necesaria la revisión posterior de un humano para garantizar la calidad de la preservación? La revisión manual permite aprovechar la experiencia de los expertos, pero, dado que es un proceso caro, quizás deba reservarse para aquellos sitios especialmente relevantes.

5. Conclusiones

Las bibliotecas deben revisar algunas de sus responsabilidades como consecuencia de la utilización de documentos electrónicos, procesos automáticos y medios de difusión sin controles de calidad como la Web. En este trabajo se han estudiado cuatro aspectos: el acceso a la información, la búsqueda de información, la calidad, y la perdurabilidad. Se ha analizado cómo les afectan los aspectos mencionados, se han presentado las propuestas que se están ofreciendo actualmente, y se han planteado cuestiones que deben ser consideradas a la hora de buscar nuevas soluciones.

Los sistemas de libre acceso resultan más baratos de mantener, ya que no es necesario invertir en la puesta en práctica de políticas de restricción de acceso. Por otro lado, también se debe considerar que los autores miden su prestigio en función del de la revista en la que publican sus trabajos (aunque esto no es inamovible). La situación actual se puede resumir en dos líneas que se enfrentan. Por un lado, hay una tendencia *comercial*, cuyo objetivo es obtener un beneficio económico de la publicación, que considera el acceso abierto como una amenaza y que utiliza su poder para presionar a su favor a la hora de legislar estos aspectos. Por el otro, un interés totalmente diferente es el de la comunidad científica y académica,

que pretende *difundir* conocimiento, que sea accesible al máximo número posible de personas. En el futuro lo más probable es que ambas tendencias coexistan en las bibliotecas digitales, que podamos acceder a un número de recursos de forma gratuita, pero que debamos pagar por ciertos procesos encaminados a garantizar la calidad de la información que obtenemos.

La tecnología ofrece soluciones automáticas que pueden facilitar la implementación de ciertas tareas, inspiradas en los métodos de trabajo de los expertos. Pero también plantea nuevos interrogantes que necesitan respuestas. Es en dar respuesta a estas preguntas —allí donde los procesos automáticos no pueden tomar decisiones—, donde los profesionales del futuro deberán hacer mayores esfuerzos.

6. Notas

(1) URL: <http://www.rfc-editor.org/overview.html>.

(2) Un ejemplo es la conferencia "XML Europe 2002" (URL: <http://www.xmleurope.com/2002>).

(3) URL: <http://gdbwww.gdb.org>.

(4) URL: <http://www.w3c.org>. Dato obtenido en noviembre de 2002, de la sala de prensa del portal. URL: < http://www.universia.es/saldeprensa/faq.html>.

(6) URL: <http://www.netLibrary.com>.

(7) URL: <http://www.adobe.com/products/ebookreader/>.

(8) Kahle, Brewster and Rod Hewitt. "Lending Library Format (.llf), " Version 1. 0. 5, February 22, 2001. URL: <http://www.archive.org/llf.html>.

(9) MPEG 1 (Moving Picture Experts Group 1), audio layer 3.

(10) Puede encontrarse más información en la URL: <http://eon.law.harvard.edu/mp3>. Véanse las referencias DC2001/29/CE y DMCA.

(12) Según (Arms, 2001).

(13) Association of Computer Machinery (ACM).

(14) Robot es el nombre genérico. También se les llama *spiders*, *crawlers* y *wanderers*.

7. Anexo 1. Cálculo de la exhaustividad y precisión en un ejemplo

Se dispone de una colección de 1000 documentos, sobre la que se realiza la consulta "Propiedad intelectual". Entre los documentos de la colección:

a) 80 contienen el término "propiedad intelectual" y tratan de este tema,

b) 40 contienen el término "copyright" y tratan sobre él,

c) 20 contienen el término "propiedad intelectual", pero en realidad no tratan sobre este tema (son documentos con reseñas sobre los derechos de propiedad intelectual que les afectan).

Suponiendo que la herramienta de búsqueda busca en texto completo:

— el total de documentos recuperados es a) + c) (100 documentos)

— la cantidad total de documentos relevantes disponibles en la colección es a) + b) (120 documentos).

Por tanto los valores de los dos índices son:

— Exhaustividad = 80/100 = 0,8

— Precisión = 80/120= 0,66

8. Agradecimientos

La autora agradece al Dr. Dámaso Javier Vicente Blanco, del departamento de Derecho Internacional Privado de la Universidad de Valladolid (España), su valiosa ayuda en la preparación de las cuestiones relacionadas con los aspectos legales. También da las gracias a la Dra. Judith Licea de Arenas (Universidad Nacional Autonóma de México), por su orientación en la localización de referencias sobre el proceso de revisión editorial que han mejorado este trabajo.

9. Referencias

Arms, W. Y. (2001). Digital Libraries Architectures and Open Access to Digital Libraries. // Tutorial. First DELOS International Summer School on Digital Library Technologies. Pisa, Italia (julio 2001).

Arms, W. Y. (1999). Open Access, Subsequent Use, and MP3. // D-Lib Magazine 5:2 (febrero 1999).

Brin, Sergey; Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. // Computer Networks and ISDN Systems. 30:1-7 (1998) 107-117.

Digital Millennium Copyright Act // U. S. Copyright Office Web Site. URL: <http://lcWeb.loc.gov/copyright/>

Directiva Comunitaria sobre derechos de autor en la sociedad de la información, Directiva 2001/29/CE relativa a la armonización de determinados aspectos de los derechos de autor y derechos afines a los derechos de autor en la Sociedad de la Información, de 22 de mayo de 2001

Kahle, Brewster; Prelinger, Rick; Jackson, Mary E. (2001). Public Access to Digital Material. // D-Lib Magazine. 7:10 (octubre 2001).

Lawrence, Steve; Giles, Lee; Bollacker, Kurt (1999). Digital Libraries and Autonomous Citation Indexing. // IEEE Computer. 32:6 (1999) 67-71.

PADI (Preserving Access to Digital Information). // URL: <http://www.nla.gov.au/padi/topics/71.html>

Weller, C. A. Editorial Peer Review: Its Strengths and Weaknesses. // Information Today Inc., 2001.