El profesional de la información
ISSN 1386-6710
Revista internacional científica y profesional
Indexada por ISI y Scopus

Julio 1998

Archivar internet

Por **Jordi Serra i Serra**

Resumen: *Cada día millones de usuarios vuelcan sus conocimientos y experiencias en internet. Este flujo imparable de información, verdadera memoria histórica del presente, no está siendo objeto de ningún tratamiento específico, y al finalizar su vigencia desaparece o es sustituido sin más. Este artículo reflexiona sobre los peligros que corre la documentación que circula por internet desde un punto de vista archivístico, y sobre las dificultades técnicas que entraña su conservación. A continuación analiza los principales proyectos que se han puesto en marcha con la finalidad de acumular un fondo documental de internet para su conservación permanente.*

Palabras clave: *Internet, Archivos, Documentos electrónicos, URL.*

Abstract: *Every day millions of people pour their knowledge and experiences into internet. This continuous information flow, that truly represents the historical memory of the present, is not being specifically managed, and eventually disappears or is simply replaced when its "usefulness" is considered ended. This article reflects upon the risks faced by the body of documents circulating on the Net, from an archival point of view, and upon the technical difficulties involved in conserving it. The article then analyzes the principal projects that have begun, whose goals are to collect and permanently preserve the documental heritage available on the internet.*

Keywords: *Internet, Archives, Electronic records, URL.*

La Red va camino de ser la memoria histórica del presente, el lugar donde se van a encontrar las opiniones y las investigaciones de la actualidad.

El reciente escándalo del presidente **Bill Clinton** ha servido para propulsar el papel de internet como medio de comunicación. Muchas de las informaciones más relevantes del caso Lewinsky aparecieron y circularon por internet mucho antes de su publicación en la prensa escrita. Ningún caso anterior había obligado a los periodistas a consultar tanto la Red para descubrir nuevas revelaciones.

Hechos como éste nos dan la medida de la trascendencia de los datos que circulan por internet. Pero, hasta el momento, no han conseguido despertar la conciencia de conservar, con criterios diversos, el inmenso legado que se acumula en la Red día a día. De alguna manera, un cambio de soporte ha derivado en un cambio de valor. Los mismos documentos que antes se consideraban de obligada conservación, al estar en formato digital se ven como menos importantes y como susceptibles de eliminación una vez finalizado su período de utilización. Es urgente tomar alguna medida antes de que se hayan destruido años de materiales y de información.

La dirección etérea

Desde el primer momento ha sido obvia la necesidad de documentar adecuadamente los documentos de internet, y en especial las páginas web, a fin de facilitar su localización y su recuperación. A este fin se ha aplicado la dilatada experiencia acumulada por los profesionales de la información en la descripción de elementos de información.

Las páginas web de más entidad han empezado ya a incorporar elementos de autodescripción, con el fin de facilitar a los robots de búsqueda la creación de los ficheros invertidos. Estos elementos, conocidos como meta-etiquetas, permiten describir la estructura lógica (*DTD*) y el contenido de los

documentos escritos en lenguaje *sgml* (*ISO 8879*). El lenguaje html, derivado del *sgml*, permite incluir en la cabecera de los documentos estas meta-etiquetas, que indican datos como el formato, el título o las palabras clave del documento.

A fin de normalizar el contenido de las meta-etiquetas, *Oclc* ha elaborado una propuesta de estándar de elementos descriptivos, el *Dublin Core* (ver IWE vol. 6, n. 11, noviembre de 1997, pp. 14-16), que consta de 15 campos de información mínimos, y de unas normas que permiten su ampliación.

El papel jugado por las meta-etiquetas se verá incrementado sin duda por la utilización del nuevo lenguaje *XML* (versión abreviada del *sgml*), que, al ser sensible al contexto, permite la personalización de la presentación de un documento a partir de su meta-descripción. La singularidad de los recursos electrónicos ha centrado el interés en su detallada descripción.

La norma más utilizada actualmente, la *ISO 690-2*, describe y ubica los documentos electrónicos en Red, a la vez que permite indicar la vigencia de las citas mediante la fecha de la última consulta. La norma internacional de descripción archivística *Isad-g* no contempla todavía la descripción de documentos electrónicos.

Todos los mecanismos descriptivos anteriores adolecen de un problema común: la única forma de citar la ubicación de una página web es, por el momento, su *URL*, la dirección electrónica, producto del sistema de referencia único de internet. Como es sabido, la asignación de direcciones *IP* y *DNS* se centraliza en la *Iana* (*Internet assigned numbers authority*), que la delega en diversos registros regionales, a escala continental para las direcciones *IP*, y a escala nacional para los nombres de dominio.

Es paradójico que los mismos investigadores que completan sus estudios con detalladísimas citas bibliográficas, se vean obligados a citar los documentos de internet únicamente por su *URL*. Todos compartimos este error, porque hasta el momento no tenemos otra dirección con la que referenciar una página web.

Con la informatización de la sociedad, la capacidad de documentar cualquier actividad humana que comporte un intercambio de información se ha multiplicado.

Ésta es la causa de que el horizonte internet sea tremendamente cambiante. Se calcula que la vida media de una web es de 44 a 75 días, y que casi un 30% de las webs cambian cada 10 días. Cuando una página web cambia de servidor recibe una nueva *URL*. Lo mismo sucede cuando se crean nuevas versiones de la página, por razones de idioma, de presentación (frames/no frames) o de formato (un mismo documento en html, pdf, ascii...). El problema radica en que, si el webmaster no se ha preocupado de notificarlo, o si los buscadores no han actualizado sus índices, esta página se puede haber perdido para muchos de sus "clientes", los cuales en su lugar encontrarán el ya típico "*Error 404. Object not found*".

Proyectos de *Oclc*

El *Online Computer Library Center, Inc.* ha puesto su experiencia al servicio de la comunidad internauta para encontrar una solución a este problema. Ya en 1991 llevó a cabo un proyecto de catalogación de documentos electrónicos que puso de relieve la adaptabilidad del formato *Marc* (aplicando las *Aacr2*) a la catalogación de dichos documentos. Hizo falta añadir un campo adicional (el campo 856: "Electronic Location and Access") a fin de relacionar el documento digital con su fuente online. Esta experiencia se amplió, en 1994, con el proyecto *Intercat*, dando origen a la base de datos *NetFirst*

http://purl.oclc.org/net/InterCat

Estas experiencias demostraron que unos mismos estándares de catalogación permitían integrar los nuevos documentos digitales de internet en los catálogos existentes. Pero también pusieron en evidencia que uno de los campos descriptivos, el campo 856, contenía una información que cambiaba frecuentemente y de forma imprevisible: la *URL* del documento. Esto ha llevado a *Oclc* a crear su propia solución: asignar a cada documento catalogado una dirección conocida como *Persistent uniform resource locator* (*Purl*), a través de un servidor de *Purl*, que redirecciona cualquier link a la

ubicación correcta

http://purl.oclc.org/

Este sistema utiliza el catálogo de *Oclc* como referencia de direcciones de páginas web, creando un nuevo registro *Purl* para cada documento catalogado, en el que el campo 856 contiene la *URL* permanentemente actualizada. De esta manera, cualquier cambio en la dirección de una página web sólo debe ser actualizado en la base de datos de *Purl*, y los links que contengan la dirección *Purl* siempre remitirán a la hoja correcta. *Oclc* se ha comprometido a actualizar los índices de todos los motores de búsqueda del mundo con las correspondientes direcciones *Purl*.

Simultáneamente, *Oclc* participa con la *Internet Engineering Task Force* (*Ietf*) en el desarrollo de los estándares *Uniform Resource Name* (*URN*). La utilización de *Purl* se considera un paso previo, a la espera de que las *URNs* sean una parte integrante de la arquitectura de información de internet.

Los documentos informáticos

Internet es el reino absoluto de los documentos informáticos: todos los formatos cohabitan en un mar de servidores entrelazados. De esta manera, la problemática inherente a los documentos informáticos se sitúa también en un primerísimo plano. ¿En qué se resume esta problemática?

En primer lugar, el lazo forma-contenido es extraordinariamente débil, atacado permanentemente por la constante evolución del software y del hardware. La vulnerabilidad de un documento aumenta en la medida que está compuesto por un mayor número de objetos en distintos formatos. Además, se trata de documentos eminentemente modificables, de lo que se resienten las garantías legales.

Por otra parte, la facilidad en la duplicación y distribución de los documentos es un atentado al principio de originalidad y de procedencia (una copia digital es, en realidad, un original duplicado).

Finalmente, si queremos fijar los documentos físicamente, disponemos de unos soportes de corta vida. Esto obliga a unas siempre costosas políticas de migración (cada 5 ó 10 años), con la pérdida de datos que cada transición comporta.

En el campo de la preservación de los documentos en soporte electrónico están en marcha diversos proyectos a nivel internacional. Uno de los más destacados es el realizado por la *Universidad de British Columbia* en colaboración con la *Records Management Task Force* del *Departamento de Defensa* de EUA:

http://www.slais.ubc.ca/users/duranti/

Incide en la recogida de los metadatos suficientes para asegurar la integridad y la conservación de los documentos en soporte electrónico. Asimismo, este tema fue objeto de debate en el primer foro sobre documentos electrónicos celebrado en Bruselas en 1996 (ver actas del *DLM-Forum on electronic records* y *Guidelines on best practices for using electronic information*, Insar, suplemento III, 1997).

El archivo digital

Las páginas web se asemejan al río de **Heráclito**: aunque la *URL* se mantenga constante durante un tiempo, difícilmente podremos encontrar en una página el mismo contenido que vimos cuando la visitamos por última vez. La elevada frecuencia de actualización de un medio tan dinámico como internet puede volatilizar cualquier información antes incluso de haber hecho mención de ella. Este hecho hace patente la necesidad de archivar (en toda la amplitud del concepto) los contenidos de internet.

El principio archivístico de procedencia es el criterio que sirve para agrupar una colección de documentos formando un fondo, que bajo ningún concepto se puede desmantelar. La integridad del fondo es la garantía de la conservación del contexto de los documentos. Toda institución que utilice internet para sus actividades da origen a un fondo documental, cuya conservación y mantenimiento es responsabilidad de la propia institución generadora (si ésta es privada) o de las diferentes administraciones (si es de titularidad pública).

Sin embargo, constatamos que no es frecuente la creación de los correspondientes archivos digitales. Este desinterés es, sin duda, resultado de una falta de conciencia de que los documentos generados por una organización son, sea cual sea su soporte, documentos con valores

administrativos, afectados por unos períodos de conservación y por unas especificaciones de acceso. Indudablemente no todos los documentos que circulan por internet tienen un valor intrínseco. Pero desde el punto de vista archivístico, el valor de los documentos no proviene únicamente de su contenido sino también de su contexto y de la relación con la institución que los ha generado.

Hasta ahora ha predominado la utilización de internet como medio de difusión por encima de su utilización como medio de negocio, por lo que circulaban más documentos cognitivos que administrativos. Precisamente por esto los bibliotecarios y los documentalistas se han visto obligados, mucho antes que los archiveros, a entrar en el universo internet. Pero esta tendencia se va invirtiendo: cada vez circularán más documentos con valor administrativo.

El concepto de "comunidades virtuales", que internet ha creado, está propiciando la aparición de "instituciones virtuales". Si partimos de que la necesidad de crear y mantener un archivo proviene del funcionamiento de una institución, es evidente que también se deberán crear "archivos virtuales" a fin de gestionar la documentación generada por estas nuevas instituciones.

No es nada fácil, para una organización, crear y mantener un archivo digital. Imaginemos una empresa pública que disponga de una intranet corporativa y que utilice internet tanto para la difusión como para las transacciones con los clientes. El primer paso será asumir que la gestión de los documentos informáticos es competencia de los profesionales de la gestión documental, hecho poco claro en la mayoría de los casos. Después habrá que rediseñar los circuitos de información, a fin de que garanticen que se documentan adecuadamente todas las acciones.

Finalmente, deberá disponer de los recursos técnicos para mantener un archivo en soporte digital, y poder garantizar la autenticidad, la integridad y la conservación de los documentos, teniendo en cuenta que mantener y facilitar el acceso a este hipotético archivo digital implica un coste permanente y creciente.

¿Quién se encarga del archivo "global"?

Hay organizaciones que no tienen capacidad (ni quizás voluntad) de aceptar la responsabilidad de crear y mantener su archivo digital. De la misma manera, hay numerosos documentos en la Red totalmente públicos, gestionados por particulares, de los que en principio nadie se debe responsabilizar. En este caso, ¿debemos dar por perdido todo este material? Existen proyectos que representan una alternativa real a este problema.

En principio se puede pensar que la mayoría de buscadores (*AltaVista*, *Yahoo*, *Lycos*, etc.) ya realizan esta función. Sin embargo, no todo lo que podemos ver a través de internet puede ser recogido e indexado sin problemas. Los buscadores más potentes apenas indexan la mitad de los muchos millones de páginas web que existen en este momento. Montones de documentos que circulan por la Red quedan por recoger, protegidos detrás de formularios de registro, *query boxes*, *firewalls* de intranets o software de exclusión de robots. El contenido de ficheros en formatos no textuales (*pdf*, imágenes *bmp*, *jpg*, etc.) no se incluye en los índices generales, como tampoco el contenido de las bases de datos que se ofrecen online por internet, aunque los resultados de las consultas interactivas se muestren en pantalla a través de un formulario en html. Por esta razón, los buscadores masivos, orientados a la localización y no a la conservación, no satisfacen esta necesidad de archivo.

Un proyecto que sí cuenta claramente entre sus objetivos el de recoger la memoria histórica que se va acumulando en internet es *Internet Archive*

http://www.archive.org

Brewster Kahle, creador del sistema *Wais*, ha puesto en marcha un gran archivo digital de internet. El método, en este caso, está siendo objeto de una cierta polémica. La conveniencia de aplicar los filtros de selección en el momento de la recopilación de los datos, a fin de almacenar únicamente la información relevante, y no "a posteriori", como practica *Internet Archive*, lleva a cuestionar este proyecto. Un archivo universal es fruto de una óptica que considera todo el conocimiento como un continuo, que se puede clasificar según un mismo criterio (normalmente temático).

Desde una perspectiva archivística nunca se puede plantear un proyecto en estos términos, porque atenta a la esencia misma del principio de

procedencia. De todos modos, la situación actual coloca al profesional de la información en una tremenda disyuntiva, pues aunque un proyecto basado en "recogerlo todo ahora y seleccionarlo después" es metodológicamente cuestionable y económicamente inviable, es notoria la necesidad de tomar medidas para corregir una situación acuciante. En previsión de su futura evaluación, en *Internet Archive* los datos se almacenan con el máximo posible de metadatos.

El alcance de los buscadores de *Internet Archive* comprende tanto el World Wide Web como los grupos de noticias de Usenet y los servidores ftp y gopher. De manera similar a como se construyen los índices de *AltaVista*, los robots buscadores de *Internet Archive* rastrean la Red y obtienen copias de todos los documentos accesibles públicamente que localizan. Con una frecuencia de visita bimensual, estos buscadores van acumulando en los servidores de *Internet Archive* unos 1,5 terabytes ($1,5 \times 10^{12}$ bytes) de datos cada mes.

Para difundir y rentabilizar el proyecto, desde el primer momento se ha ofrecido servicio público. A través del software *Alexa* (ver IWE v. 7, n. 4, abril de 1998, p. 20), que funciona de forma totalmente integrada con cualquier navegador, se puede recuperar del archivo de internet una página desaparecida cuando un link da el mensaje *Error 404. Object not found*.

Además, es posible obtener información acerca de la web que se consulta, su localización, número de páginas, velocidad del servidor y popularidad (basada en la frecuencia de visitas). Actualmente, este servicio de acceso tiene todavía una funcionalidad limitada.

El programa *Alexa* se puede obtener vía ftp en:

http://www.alexa.com

Hay otros proyectos que, aun con objetivos distintos, tienen en común el hecho de que están acumulando un fondo de documentación, recopilada de internet, con la vocación de conservarla permanentemente. En la situación actual, cualquier depósito de datos acumulado con cualquier finalidad adquiere ya un cierto valor "histórico". Citaré algunos ejemplos:

A Business Compass

http://abcompass.com

Es un sistema que supera las barreras de registro, lo que le permite indexar regularmente cerca de 1.200 lugares de interés para el mundo empresarial. Como sistema de archivo, almacena todas las publicaciones exclusivamente electrónicas del ámbito de los negocios, las clasifica y, previa suscripción, facilita acceso a un resumen y al link de la publicación original.

Cuando un artículo ya no está disponible online, *A Business Compass* negocia con el propietario el permiso para usar su copia almacenada para acceso público. Este proyecto tiene la voluntad de ir acumulando un fondo de publicaciones para su conservación permanente. El servicio de publicaciones se complementa con un rastreador de grupos de noticias que almacena los contenidos y genera unos resúmenes automáticos, utilidad que puede representar un buen sistema de almacenamiento para los efímeros mensajes de news.

eWatch

http://www.ewatch.com

Fundado en 1995, comenzó recogiendo el texto completo de los mensajes de grupos de news y listas de distribución de *CompuServe*, *Prodigy*, *America OnLine* o *Microsoft Network*. Actualmente el sistema procesa cerca de 250.000 mensajes al día.

ECO

Oclc ha puesto en marcha el servicio *Electronics Collections Online* (*ECO*), que, previa suscripción, mantiene una colección de revistas electrónicas, obteniendo la fuente directamente de las editoriales, y garantizando la conservación, el acceso continuado y la migración a sistemas tecnológicos más modernos (v. IWE, vol. 7, n. 4, pp. 14-16).

Siglas empleadas

aacr2: *anglo-american cataloguing rules version 2*

ascii: *american standard code for information interchange*

bmp: *bitmap picture* (.extensión)

DNS: *domain naming system*

DTD: *document type definition*

IP: *Internet protocol*

Isad(g): *general international standard archival description*

ISO: *International Standards Organization*

jpg: *jpeg, join photographic experts group* (.extensión)

marc: *machine readable cataloguing*

pdf: *portable document format* (.extensión)

Sgml: *standard generalized markup language*

URL: *uniform resource locator*

wais: *wide area information server*

Bibliografía

Aguayo, Pablo. "Documentar las páginas web", participación en el foro electrónico *Extra!-net*, 25 de febrero de 1998:

http://www.extra-net.net/forum/missatges/138.html.

Estivill, Assumpció; **Urbano, Cristóbal**. "Cómo citar recursos electrónicos". En *Information World en Español*, vol. 6, n. 9, septiembre 1997, pp. 16-26.

http://www.ub.es/div5/biblio/citae-e.htm

Feldman, Susan E. "It was here a minute ago!: archiving the net":

http://www.infotoday.com/searcher/oct/story4.htm

Serra, Jordi. "Internet: memoria de la humanidad", en *Net-conexión*. n. 28, marzo 1998, pp. 43-47.

Jordi Serra i Serra. *Arxiu central del Departament de Cultura de la Generalitat de Catalunya.*

jordis@bcnet.upc.es

Enlace del artículo:
http://www.elprofesionaldelainformacion.com/contenidos/1998/julio/archivar_internet.html