

Internet Archive o el regreso al pasado

EL 24 DE OCTUBRE DE 2001 *Internet Archive*, una organización que mantiene una colección de páginas web y materiales digitales, lanzó la *Wayback Machine*¹ (“máquina de regreso al pasado”), el acceso abierto al propio recurso. En 1996 *Internet Archive* había empezado la recopilación de estos documentos para su preservación, y 5 años después ha empezado a proporcionar el acceso a los 10.000 millones de páginas que ya tiene almacenadas para todo el público.

Internet Archive es una organización sin ánimo de lucro ubicada en California ideada por **Brewster Kahle**. La idea se le ocurrió después de participar en el proyecto *Thinking Machines* del *Massachusetts Institute of Technology (MIT)* donde trabajaba con supercomputadoras, ampliándola con posterioridad cuando profundizó sobre la memoria histórica a través de estudios de biblioteconomía². *Wais* (*wide area information server*) fue el siguiente proyecto, un sistema de publicación en red que vendió a *America Online*. Entonces **Kahle** lanzó *Alexa*³, un programa gratuito que se integra en los navegadores para obtener información sobre las páginas visitadas y que vendió a *Amazon* el año 2000, aunque continúa en la directiva de la empresa. Desde 1996 todo lo que rastrea *Alexa* en la Red se almacena en *Internet Archive*.

Kahle le puso el nombre de *Alexa* pensando en el objetivo de la antigua *Biblioteca de Alejandría*: recoger todo el material que se publicaba. Con la tecnología existente y con la capacidad de memoria actual, cree que es posible crear una *Alejandría* electrónica. Afirma que todos los documentos que to-

das las personas del mundo puedan teclear a lo largo de su vida se pueden archivar, porque ahora ya no se trata de una cantidad de memoria que dé miedo⁴. Los costos del esfuerzo que de momento está haciendo *Internet Archive* rondan los 2 millones de US\$ anuales, según la misma empresa.

¿Qué se almacena en *Internet Archive*?

Decidir qué es lo que se debe preservar es una dificultad inicial con la cual se encuentran todas las bibliotecas y archivos. Pero no es sólo un problema conceptual sino que también el coste y las complejidades de las soluciones tecnológicas requeridas influyen en la decisión. *Internet Archive* tiene como objetivo almacenar la Red entera y hacerlo diariamente, pero por el momento sólo cubre textos en ascii y páginas web públicas, mientras que no recoge ni imágenes ni otros objetos multimedia.

La web pública se define como la que se tiene acceso a través del protocolo web (http) y sin necesidad de contraseña. Actualmente, según *Oclc*, se puede considerar que es el 36% del total de la web y que la mitad de esta información la proporcionan organizaciones e individuos norteamericanos, un 5%

de proviene de Alemania y un 4% de canadienses y japoneses⁵.

«Según *Alexa* en internet cada día se añaden 1,5 millones de páginas y cada semana desaparece un 1% de ellas»

Alexa se limita a archivar los contenidos accesibles de internet y está dispuesta a retirar el material si quien disfruta del copyright así lo requiere. Sin embargo no se avisa al propietario de la página de sus actividades, tal como lo hace cualquier motor de búsqueda. Según **Lawrence Lessing**, experto en propiedad intelectual del ciberespacio, el proyecto actual de **Kahle** es una ofensa criminal potencial, y pronto aquellos que disponen del copyright van a llevarlo a los tribunales⁶. Sin embargo el fundador del proyecto no se detiene en las páginas web, está pensando en añadir otro tipo de material: libros, películas así como otras creaciones, esperando que los propietarios de los derechos de explotación estén dispuestos a participar en él.

Internet Archive no almacena páginas de sitios usenet ni ftp y no recoge mensajes de e-mail enviados a listas de distribución. Seleccionar

qué mantener se ha dejado en manos de consideraciones técnicas y comerciales. Según **Toby Burrows**⁷ el rol del bibliotecario en la selección durante la era pre-digital ya no es necesario para dejarse a merced de:

—Preservar lo que sea más fácil hacerlo. Nos podemos encontrar que, justamente lo que es más sencillo, tiene menos interés cultural.

—Permitir al mercado que decida qué seleccionar. Se conservará lo que tenga más demanda considerando que, al ocurrir esto, es lo más importante.

—El material involucrado, lo que genera preguntas tales como ¿son las páginas web más importantes de preservar que las usenet?

—Habrà de tenerse en cuenta aquello que esté en peligro de extinción. Estamos en una situación de emergencia, las páginas web desaparecen por segundos y se deben tomar las decisiones en poco tiempo.

¿Qué hace *Internet Archive*?

Utiliza el programa *Alexa* para la selección de las webs y funciona de la siguiente manera: investiga cómo los usuarios se mueven de una página a otra, permite que voten los sitios que encuentran más interesantes y proporciona una estadística de la página web indicando cuántos usuarios la han visitado, quién tiene su dominio, cuántos enlaces se dirigen a ella, cuántas páginas tiene la sede web y la frecuencia con la que se actualiza⁸. Aparte de esta información, también utiliza un robot que rastrea la web cada 6 u 8 semanas y que incrementa potencialmente el catálogo⁹.

Según *Alexa* en internet cada día se añaden 1,5 millones de páginas y cada semana desaparece un 1% de ellas. Actualmente la bdd tiene 10.000 millones de páginas web, 100 terabytes de memoria y crece 12 terabytes al mes. Si se di-

The screenshot shows the Internet Archive website. At the top, it says "The Internet Archive: Building an 'Internet Library'". Below this, there are navigation links for "Internet", "Movies", "Arpanet", and "About the Archive". The main content area features several sections: "Announcements" with links to various news items, "The Internet Archive is building a digital library" with a description of their mission, and a large section for the "Wayback Machine" which includes a search bar and a "Take Me Back!" button. Below the search bar, there are "Special Wayback Collections" for "September 11, 2001", "Web Pioneers", and "Election 2000". The page also lists "Archive Users" and "Archive Donors".

SISTEMA SABINI

Soluciones integrales para la Automatización de Bibliotecas y Centros de Documentación

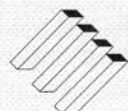
Desarrollado íntegramente en castellano con más de 15 años de experiencia en España y Latinoamérica, atiende las necesidades de todo tipo y tamaño de bibliotecas y centros de información

MÓDULOS:

- **Adquisiciones** Trámite de pedidos
Gestión de gastos y proveedores
- **Catalogación** Catalogación de todos los materiales bibliográficos
Integración de información bibliográfica
- **Terminología** Ficheros de autoridades
Tesauro multilingüe
- **Consultas** Lenguaje de Comandos
Acceso público en línea (CAPEL)
Acceso a través de WWW
Protocolo Z39.50
- **Circulación** Gestión de préstamos
Fichero de usuarios
- **Estadísticas** De proceso y circulación

Otros servicios:

- Instalación y soporte técnico del sistema SABINI
- Asesoría en Automatización de Bibliotecas y Centros de Documentación
- Procesamiento de material bibliográfico
- Instalación de catálogos en Internet



SABINI
Automatización
de Bibliotecas

C/ Amor de Dios, nº 1
Telf.: +34 91 4292551
Fax: +34 91 4292610
28014 MADRID
e-mail: sabini@sabini.com



DOCUMENTACIÓN
BIBLIOTECAS E
INFORMÁTICA, S.A.

Av. Diez Canseco 236 of. 602 Lima 18
Telefax: (511) 446-0815 e-mail: sabini@terra.com.pe

gitalizaran todos los libros de la *Biblioteca del Congreso*, una de las más grandes del mundo con 100 millones de libros, sólo ocuparía 20 terabytes de espacio de memoria¹⁰. La información es mantenida en una red de 300 PCs, 20 de los cuales se destinan a recibir consultas y a redirigirlas a las máquinas que tienen la información requerida. *Alexa* utiliza xml en vez de html para proporcionar las respuestas¹¹.

«La Biblioteca del Congreso de los EUA recibió de *Alexa Internet* hace ya 4 años una copia de los archivos de las páginas web que por entonces tenía *Internet Archive*: 2 terabytes de contenido»

De momento todos los ordenadores están ubicados en el mismo sitio: el edificio *Presidio Park* de San Francisco, aunque para evitar lo que le pasó a la biblioteca de Alejandría, **Kahle** ya ha pensado en hacer diferentes copias del material y distribuirlas por toda la geografía. Aparte de asegurarse hipotéticos accidentes, *Internet Archive* también ha previsto realizar migraciones cada 10 años para asegurar la preservación y evitar así la degradación del medio de almacenamiento¹².

En este proyecto ya participan diversas instituciones, entre ellas la *Biblioteca del Congreso* de los EUA, que recibió de *Alexa Internet* hace ya 4 años una copia de los archivos de las páginas web que por entonces tenía *Internet Archive*: 2 terabytes de contenido. Con este regalo **Kahle** pretendía estimular a las bibliotecas a que aceptaran la responsabilidad de preservar el conocimiento que se encuentra en este nuevo medio que es internet, tal como ya hacían con el que se en-

cuentra en papel. La donación se incorporó en el programa de digitalización de la colección de la *LoC* y se aceptó con dos propósitos: preservarlo en archivo y utilizarlo como experimento para futuras preservaciones de páginas web¹³.

Observando cómo los documentos web desaparecen, y con ellos también los materiales que describen y hacen comprender nuestra cultura, política, economía, etc., la primera reacción es almacenar. Por eso, a corto plazo, la tarea que realiza *Internet Archive* es muy valiosa, pero mientras tanto es necesario que los profesionales de la información investiguemos sobre cómo seleccionar las colecciones de materiales nacidos ya en formato digital, qué metodología de acceso es la más conveniente, etc. A medio plazo, como apunta **Susan Feldman**¹⁴, son las bibliotecas y archivos los que tienen que almacenar los materiales y tener el control sobre ellos. No son organizaciones privadas las que deben tener en sus manos el mantenimiento y el acceso de los bienes públicos, pues para ello se necesitan importantes presupuestos que apoyen estas iniciativas.

Alice Keefe, profesora de preservación digital en la *Univ. Oberta de Catalunya* y de recursos electrónicos en la *Univ. de Barcelona*, respalda esta idea: “Aunque tengan interés y capacidad para realizar la preservación, las empresas productoras o suministradoras de información no pueden ofrecer garantías de continuidad. Están sometidas a los cambios del mercado y, por tanto, sus objetivos pueden variar y ellas incluso desaparecer. Hay que fomentar la creación y gestión de centros o agrupaciones de bibliotecas capacitados para la preservación digital, dotados con los recursos adecuados”. **Keefe** añade que lo que estas iniciativas también necesitan son acuerdos so-

bre las pautas y directrices a seguir para preservar el material digital.

Lo que sí puede asegurarse es que el proyecto *Internet Archive* ya es todo un éxito. Desde que se ha dado el acceso a la bdd ha recibido más visitas de las esperadas, y es por ello que nos puede ocurrir que el servidor responda con una negativa de acceso un tanto irónica: “Access to our past will be available in the near future” (el acceso a nuestro pasado estará disponible en el futuro).

Notas

1. <http://web.archive.org>
2. **Pisani, F.** “La memoria multimedia del mundo”. En: *CiberPaís*, 2001, 15 de marzo.
3. <http://www.alexa.com>
4. **Wiggins, R.** “Digital preservation: paradox & promise”. En: *Library journal*, 2001, primavera.
5. *Oclc researchers find slowdown in web growth*, 4 oct. 2001.
<http://www.oclc.org/oclc/press/20011004a.shtm>
6. **Schwartz, J.** “A library of web pages that warms the cockles of the wired heart and beats the Library of Congress for sheer volume”. En: *New York Times*, 2001, 29 de octubre.
7. **Burrows, T.** “Preserving the past, conceptualising the future: research libraries and digital preservation”. En: *Australian academic & research libraries*, 2001, v. 31, n. 4, p. 142.
8. **Bates, M. E.** “Alexa Internet”. En: *Database*, 2001, v. 22, n. 2, p. 12.
9. **Quint, B.** “A gift of the web for the Library of Congress from Alexa Internet”. En: *Searcher*, 1998, noviembre, p. 12.
10. **Hamelick, A.** “Internet Archive launches Wayback Machine”. En: *JoDI noticeboard*, 2001.
<http://jodi.ecs.soton.ac.uk/noticeboard/wayback.html>
11. **Roberts-Wilt, S.** “Alexa bulks up to field queries from its browsers companion”. En: *Internet world*, 1998, 5 de octubre, p. 43.
12. Storage and preservation of the collections.
<http://www.archive.org/abaout/storage.html>
13. **Quint, B.**, op. cit., 1998.
14. **Feldman, S.** “It was a minute ago!: archiving on the Net”. En: *Searcher*, 1997, v. 5, n. 9, p. 52.

Núria Ferran Ferrer, Biblioteca de la UOC.
nferranf@uoc.edu