Soluciones de archivo web libres y de código abierto

Guillermo Castellano @guillearch

La práctica de archivar sitios web se remonta prácticamente al comienzo de la comercialización de internet. En 1996, Brewster Kahle fundó Internet Archive con el objetivo de "llevar la Biblioteca de Alejandría un paso más allá y hacer accesibles a todo el mundo los trabajos publicados por la humanidad", y ese mismo año fueron capturadas las primeras colecciones

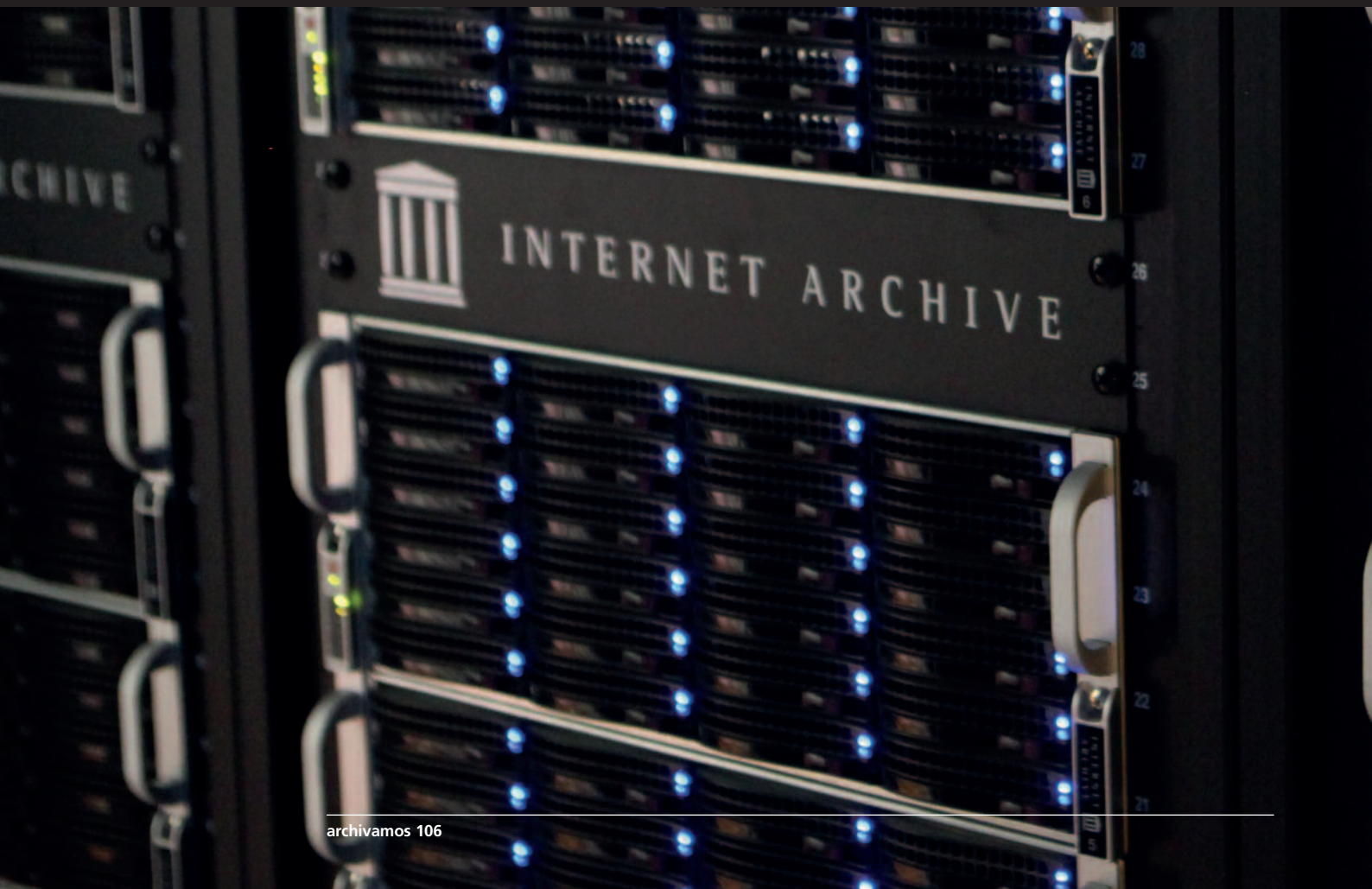
Brewster Kahle era consciente de que los sitios web representan una porción muy importante de nuestro patrimonio digital que puede perderse para siempre si no se adoptan medidas de preservación digital específicas.

El archivo web plantea un nuevo reto político y técnico para los profesionales de la información. A nivel político, los problemas más apremiantes son la censura en internet ejercida por estados y empresas privadas; la necesidad de establecer una política de conservación, ante la inviabilidad de almacenar indefinidamente todo el contenido disponible en la web; y, en tercer lugar, la falta de conciencia en la sociedad sobre el valor jurídico e histórico de la información contenida en los sitios web. A nivel técnico, esta labor tiene que asegurar la integridad del contenido y una experiencia de navegación lo más fiel posible a la original; registrar los cambios experimentados por los sitios web; y, finalmente, permitir al usuario la recuperación de la información.

Para asegurar la integridad y fidelidad de los sitios web archivados, Internet Archive desarrolló el formato ARC, del que se derivó poco después **WARC (Web ARChive)**. Lo interesante de este formato utilizado para el archivo de sitios web –definido como un estándar internacional en la ISO 28500:2009 y, más recientemente, en la ISO 28500:2017– es que permite agrupar en un único archivo todos los objetos digitales que constituyen una página web (páginas HTML, hojas de estilo CSS, líneas de código JavaScript, imágenes, archivos PDF, etc.). La ISO 28500:2009 espera que "el formato WARC se convierta en una forma estándar de estructurar, gestionar y archivar decenas de millones de recursos obtenidos de la red y otros lugares" y sea "utilizado para desarrollar aplicaciones de captura, gestión, acceso e intercambio de contenido". WARC es un formato abierto, lo que significa que sus especificaciones técnicas están publicadas y que no existe la obligación de pagar licencias para desarrollar una aplicación informática que sea compatible con él.

Internet Archive, en colaboración con las bibliotecas nacionales nórdicas, desarrolló también **Heritrix**, un rastreador web (o *web crawler*) libre y de código abierto que captura de manera remota las URL de un dominio y las convierte a formato WARC. Hasta que esta herramienta estuvo terminada en enero de 2004, Internet Archive se había nutrido de la base de datos de Alexa Internet, una empresa subsidiaria de Amazon dedicada al análisis comercial del tráfico web.

El desarrollo más conocido de Internet Archive probablemente sea **Wayback Machine**, una aplicación web que permite acceder a las URL capturadas por medio de un rastreador. Su popularidad se



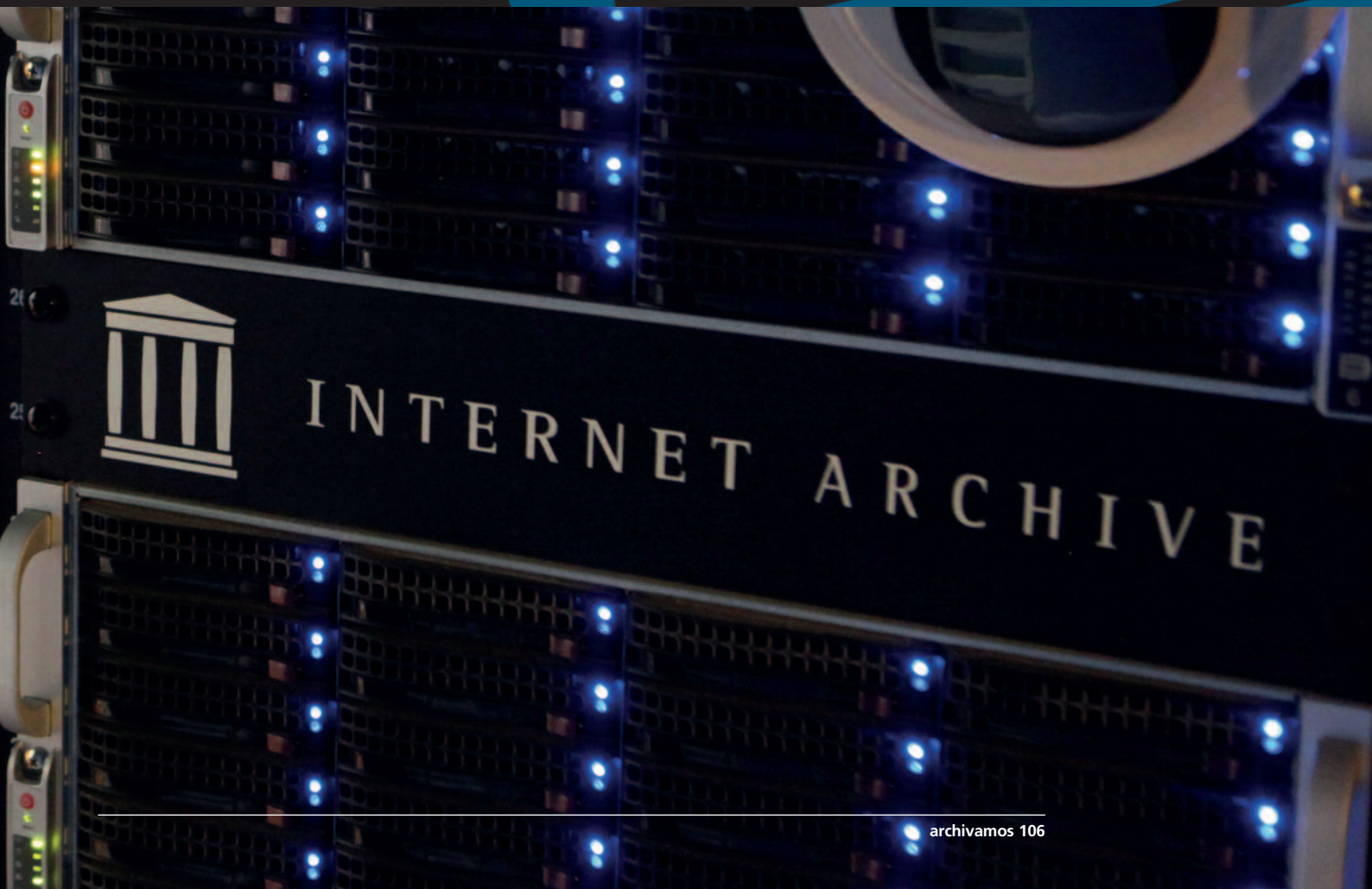
debe a que Internet Archive lo utiliza en su sitio web para dar acceso a los más de 380 mil millones de sitios que ha archivado a lo largo de las dos últimas décadas. Una de las funcionalidades destacadas de Wayback Machine es que agrupa estas capturas por fechas, de tal modo que permite al usuario ver la evolución del sitio web y el contenido que se ha ido eliminando.

Los archivos WARC también pueden ser indexados utilizando **NutchWAX (Nutch Web Archive eXtensions)** o buscadores basados en Lucene como Solr y Elasticsearch. Mientras que Wayback Machine utiliza la URL original o metadatos para recuperar el contenido del sitio web, NutchWAX, Solr y Elasticsearch –todas, aplicaciones libres y de código abierto– permiten realizar búsquedas mediante texto libre, lo que hace de ellas un complemento imprescindible para cualquier proyecto de archivo web de cierto volumen.

Este ecosistema de aplicaciones resuelve los tres principales retos de orden técnico que hemos identificado más arriba: mantener la integridad, controlar las versiones y proporcionar acceso. Aunque es posible descargar e instalar todas estas piezas por separado, existen plataformas que cubren todo el proceso de archivo web, desde la captura de contenido hasta su visualización por parte del usuario final.

Un ejemplo es **Archive-It**, un servicio de archivo web en la nube (SaaS) mantenido por Internet Archive y basado en Heritrix y Wayback Machine. El importe de la suscripción anual se calcula en base al número de sitios que se quiere archivar, el espacio que ocupan y la frecuencia con la que se va a capturar. La mayor ventaja de esta opción es que Internet Archive guarda tres copias de los sitios web archivados y asegura el almacenamiento perpetuo de los mismos; o, dicho de otra manera, los sitios webs archivados mediante Archive-It continúan en línea para siempre aunque el cliente deje de pagar la cuota.

Fuera del ámbito estadounidense, hay que destacar plataformas como PWA (Portguese Web Archive), desarrollada para las necesidades Arquivo.pt, o **NAS (NetarchiveSuite)**. Esta herramienta utiliza Heritrix para la captura y Viewerproxy o Wayback Machine para el acceso. Su desarrollo es el resultado de una historia de colaboración entre bibliotecas que comenzó en Dinamarca, donde la Biblioteca Real y la Biblioteca del Estado y la Universidad se pusieron a trabajar en 2004 en un programa para preservar y difundir los sitios web daneses. Al año siguiente lo pusieron en producción y en 2007 lo liberaron bajo licencia GPL, lo que permitió que se fueran sumando al proyecto instituciones de otros países, incluida la Biblioteca Nacional de España, que utiliza este software libre desde 2014



para la elaboración del Archivo de la Web Española.

La Biblioteca Nacional define el **Archivo de la Web Española** como "la colección formada por los sitios web (incluidos blogs, foros, documentos, imágenes, vídeos, etc.) que se recolectan con el fin de preservar el patrimonio documental español en Internet y asegurar el acceso al mismo". Este proyecto, basado a nivel técnico en NAS, ofrece colecciones selectivas (por temas como la muerte de Adolfo Suárez, la abdicación de Juan Carlos I y la proclamación de Felipe VI, los últimos procesos electorales o los atentados terroristas en Cataluña) y una captura masiva anual de todos los dominios registrados en ESNIC.

Otras instituciones y organizaciones de nuestro ámbito geográfico que utilizan estas aplicaciones y plataformas de archivo web libres y de código abierto son el Servicio de Bibliotecas del Gobierno Vasco, la Biblioteca de Cataluña y Greenpeace España.

Una de las grandes ventajas del software libre y de código abierto es que estimula el libre intercambio del conocimiento y permite a archivos y bibliotecas de diferentes partes del mundo desarrollar colaborativamente las herramientas informáticas que necesitan. Formatos como WARC, aplicaciones como Heritrix, Wayback Machine, NutchWAX y plataformas como Archive-It y NAS demuestran el enorme potencial del software libre para responder a los nuevos retos que tenemos por delante los profesionales de la información, como el archivo de los sitios web.

La fortaleza de este modelo de desarrollo reside en el dicho popular de que "cuatro ojos ven más que dos". En el mundo del software libre, esta cualidad del desarrollo colaborativo se conoce como Ley de Linus, originalmente formulada por Eric Raymond en los siguientes términos: "dada una base suficiente de desarrolladores asistentes y beta-testers, casi cualquier problema puede ser caracterizado rápidamente, y su solución ser obvia al menos para alguien". Esto se traduce en desarrollos más eficientes, en el sentido de que ofrecen un código de mayor calidad a un coste menor.

Sin embargo, creo no estar equivocado si afirmo que la apuesta de las instituciones memorísticas por el software libre no obedece solamente a los beneficios prácticos del desarrollo colaborativo. Los archivos y las bibliotecas llevan en su ADN el principio de acceso universal al conocimiento, lo que hace que sean más receptivos a trabajar con programas que respetan la libertad de las personas para ejecutar, estudiar, mejorar y difundir el código en el que están escritos.■