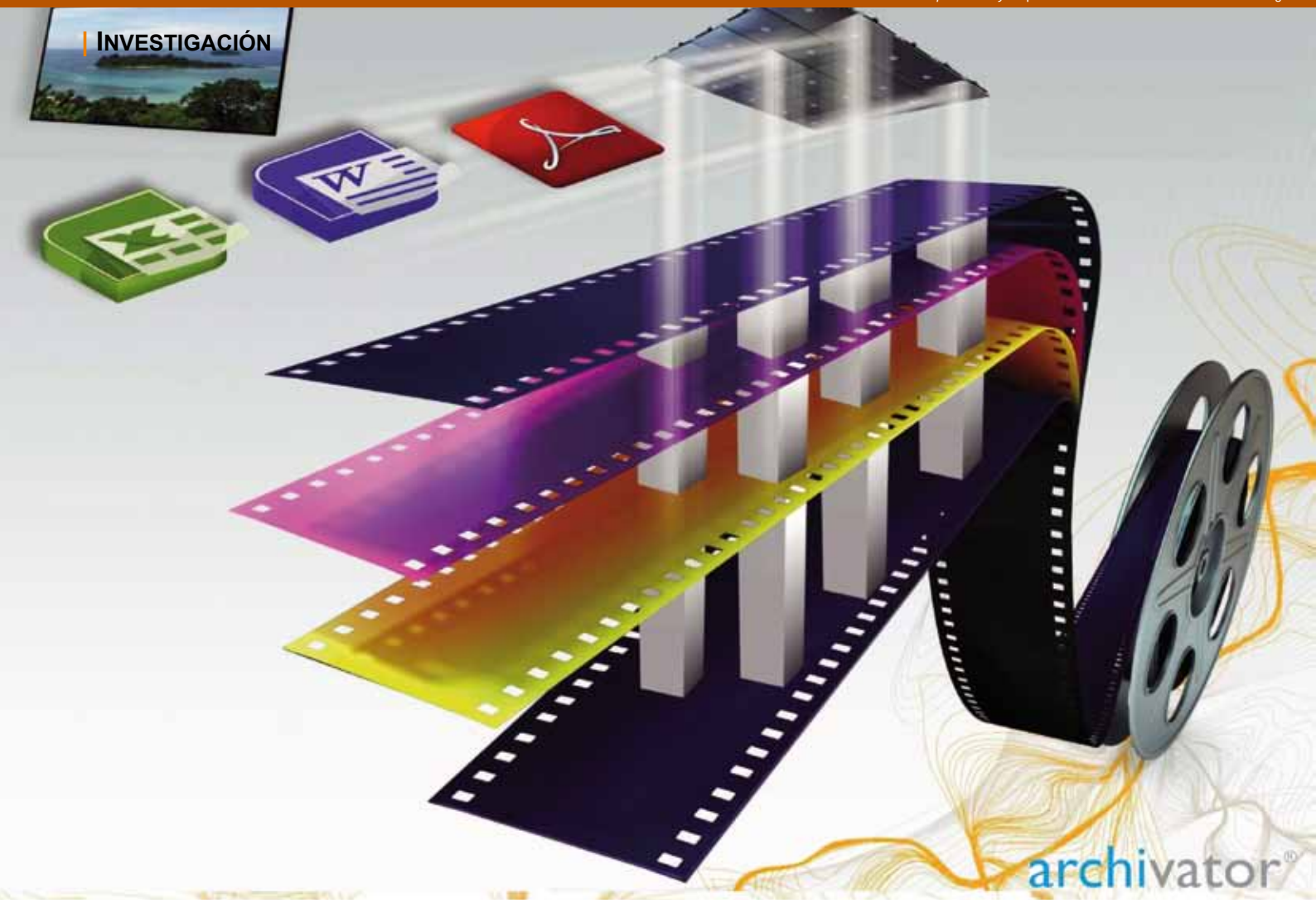


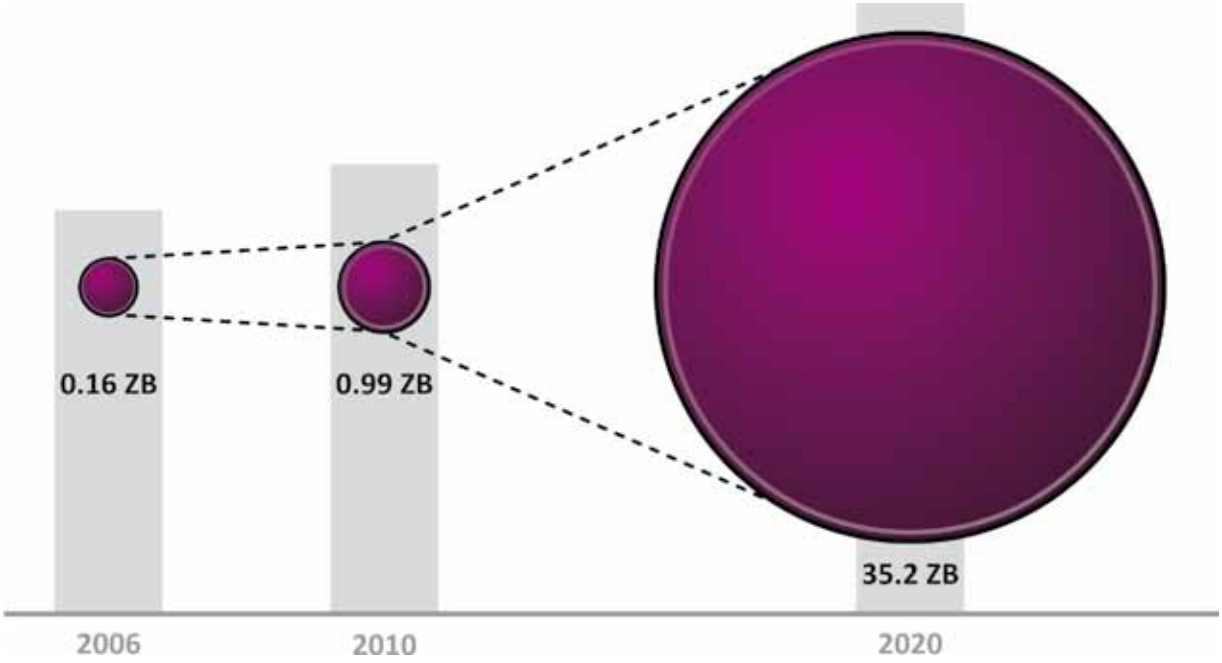
Tecnologías para el archivo de información digital a largo plazo

Las tecnologías digitales han aumentado enormemente nuestra capacidad para almacenar cantidades ingentes de información. Hoy, prácticamente cualquier tipo de información se guarda en algún dispositivo digital. Sin embargo, esta vertiginosa capacidad de almacenamiento no se ha visto acompañada de un aumento equivalente en longevidad.

> **Óscar Plata González - Emilio López Zapata** / *Catedráticos de Arquitectura y Tecnología de Computadores*
Antonio Larrosa Jiménez - Pedro Segura Plaza / *Ingenieros de I+D de Tedial*

Literalmente, nuestro planeta está saturado de información digital. A cualquier lugar que dirijamos nuestra mirada, en nuestros puestos de trabajo, en las universidades, en los museos, en las bibliotecas, en los medios de transporte, y hasta en nuestras propias casas. En todas partes se están creando datos digitales a una velocidad exponencial. Afortunadamente, el desarrollo tecnológico ha permitido el almacenamiento de esta enorme cantidad de información

Afortunadamente, el desarrollo tecnológico ha permitido el almacenamiento de esta enorme cantidad de información



En 2006 nuestro universo digital medía unos 160 exabytes, alcanzando casi 1 zettabyte en 2010, y se espera que supere los 35 en 2020

en soportes digitales de diversos tipos (discos y cintas magnéticas, CDs, DVDs, memoria flash…).

Hace apenas unas pocas décadas medíamos los datos digitales en cantidades de bytes, kilobytes y megabytes, sin embargo, en poco tiempo empezamos a utilizar las unidades de gigabytes y terabytes. En la actualidad no es difícil encontrar sistemas capaces de almacenar del orden de petabytes. Cada uno de estos términos supone un incremento de medida del orden de mil veces el término anterior, tal como se indica en la tabla. Como ejemplos reales, un terabyte de datos es equivalente a una hora de vídeo 2k (resolución de 2048 x 1080 píxeles) no comprimido, mientras que un zettabyte de datos es equivalente a la cantidad de información que una persona generaría si escribiese mensajes en Twitter continuamente durante 150 veces la edad del Universo. Como se muestra en la figura, en 2006 nuestro universo digital medía unos 160 exabytes, alcanzando casi 1 zettabyte en 2010, y se espera que supere los 35 en 2020.

Gestionar esta enormidad de datos digitales constituye un gran desafío. Por un lado, tenemos el problema de la longevidad de la información y, por otro, el de la longevidad del soporte. Con respecto a la primera cuestión, existe un riesgo creciente de pérdida de la información almacenada digitalmente debido a factores como la corrupción de datos, fallos en el dispositivo de almacenamiento, virus informáticos… Con respecto a la duración del soporte, la vida media de los dispositivos actuales es reducida, no más de 20 o 25 años. Para combatir estos problemas, actualmente se recurre a realizar tareas de migración de datos a nuevos soportes, típicamente cada tres (discos magnéticos) o cinco (cintas magnéticas) años. Sin embargo, estas políticas están sujetas a

>>

Medidas de los datos digitales	
1 bit	
1 byte	8 bits
1 kilobyte	1024 bytes
1 megabyte	1024 kilobytes
1 gigabyte	1024 megabytes
1 terabyte	1024 gigabytes
1 petabyte	1024 terabytes
1 exabyte	1024 petabytes
1 zettabyte	1024 exabytes
1 yottabyte	1024 zettabytes
1 brontobyte	1024 yottabytes
1 geobyte	1024 brontobytes
1 saganbyte	1024 geobytes
1 jotabyte	1024 saganbytes

Es imperativo encontrar una solución de archivo digital que sea duradero -unos 100 años-, seguro, robusto a la obsolescencia software y hardware, y rentable

nuevos desafíos, como la dificultad de control sobre la migración de cantidades ingentes de datos, evitando la corrupción y alteración de la información, o el problema de la obsolescencia de formatos y protocolos en los que se codifican y acceden los datos digitales.

Pero, ¿realmente hay tanta información que necesite ser almacenada durante largo tiempo? Algunos estudios indican que entre el tres y el diez por ciento de los datos digitales existentes requieren preservarse durante 25 o más años. Por tanto, es imperativo encontrar una solución de archivo digital que sea duradero (digamos, más de 100 años), seguro, fiable, robusto a la obsolescencia software/hardware y rentable.

En los últimos años, algunas iniciativas han empezado a aportar soluciones a este problema. Una de las más importantes es la SNIA (*Storage Networking Industry Association*), una asociación sin ánimo de lucro dedicada al desarrollo de tecnologías para la gestión y almacenamiento de información digital. Una de sus iniciativas es el DMF (*Data Management Forum*), centrado en los métodos para mantener la integridad de los datos, y que incluye entre sus grupos de trabajo la LTACSI (*Long Term Archive and Compliance Storage Initiative*). En 2004 crearon el 100YrATF (*100 Year Archive Task Force*), con el objetivo de establecer recomendaciones y estándares para preservar la información digital a largo plazo.

Por otro lado, el CCSDS (*Consultative Committee for Space Data Systems*), un forum internacional dedicado al desarrollo de estándares para sistemas de datos y comunicaciones espaciales, publicó en 2002 el estándar OAIS (*Open Archival Information System*), que constituye el modelo de referencia actual para preservar datos digitales a largo plazo.

Para diseñar una solución integral, esa actividad de estandarización debe verse acompañada del desarrollo de nuevos soportes de almacenamiento perdurables. Como muestra de actividad en esta línea, a mediados de 2009 se establece el consorcio industrial ARCHIVATOR con el objetivo de desarrollar una solución de archivo digital de larga duración, fiable, seguro y rentable, utilizando como soporte físico película micrográfica de alta resolución. Este tipo de soporte utiliza poliéster fotosensible, un material muy

ARCHIVATOR ofrece una solución al problema de almacenamiento de larga duración y asegura su integridad

estable que permanece inalterado por un período de siglos en condiciones óptimas de conservación.

El proyecto ARCHIVATOR pretende ofrecer una solución completa al problema del almacenamiento de larga duración de datos digitales y asegurando su integridad. Consiste en un soporte físico estable y duradero y una capa software que permite codificar los datos de forma adecuada e integrar el sistema en una infraestructura informática como si fuera un dispositivo de almacenamiento más. La organización de la capa software es similar al modelo OAIS, con dos módulos básicos encargados de la ingesta y recuperación de datos. El módulo de ingesta codifica los datos digitales en un formato especialmente adaptado a la película micrográfica, añadiendo códigos extra para la detección y recuperación de errores. Estos dos módulos están, a su vez, integrados en un gestor de procesos que automatiza los flujos de operación requeridos para almacenar y recuperar datos en el sistema.

Este gestor, el corazón del sistema, se deriva de la tecnología desarrollada por Tedial para otros sectores. Un aspecto de gran importancia en un sistema de estas características es que debe ser robusto a la obsolescencia software y hardware. Esto significa que los datos deben almacenarse en el medio físico en un formato que sea auto-contenido, auto-descrito y extensible. Un ejemplo de formato de estas características es SIRF (*Self-contained Information Retention Format*) de la SNIA, que está en desarrollo.

En el consorcio colaboran dos empresas noruegas, una alemana, una austriaca y dos españolas, siendo una de ellas la malagueña Tedial, ubicada en el PTA. Las actividades del consorcio están financiadas por el programa europeo Eureka Eurostars, y cuenta con la participación del Departamento de Arquitectura de Computadores de la Universidad de Málaga, como asesor técnico en colaboración con Tedial.

Soluciones como ARCHIVATOR tienen un mercado potencial enorme. Un sistema de este tipo es tremendamente útil para cualquier organización, institución o empresa que considere que la seguridad y la accesibilidad de información crítica a largo plazo es un factor fundamental de su éxito futuro. Tenemos ejemplos en las instituciones gubernamentales, centros de documentación, bibliotecas, museos, agencias regulatorias, agencias de seguridad, registros oficiales, compañías de seguros, bancos, centros médicos o la industria farmacéutica, entre muchos otros. ●

Los datos deben almacenarse en el medio físico en un formato que sea auto-contenido, auto-descrito y extensible