Colecciones de radio, televisión y audiovisuales en la Web: continuidad y nuevos retos

Claude Mussou
Institut national de l'Audiovisuel (INA)
Bry-sur-Marne, France

Translated by:
Alicia García Medina
Biblioteca Nacional
Madrid, Spain

Session:

148 — *Copyright law and legal deposit for audiovisual materials* — Audiovisual and multimedia with Law Libraries

Resumen:

Desde su creación el INA ha tenido como misión la de recopilar, conservar y difundir las colecciones audiovisuales en Francia. En un principio estaba pensado para favorecer la conservación de los programas radiofónicos de acuerdo con las necesidades de los profesionales, pero pronto creció y se convirtió en el gran repositorio del patrimonio audiovisual francés. De hecho, el INA, en la actualidad, es el mayor archivo de materiales audiovisuales del mundo, conservando más de 4 millones de horas de televisión y de grabaciones radiofónicas que se remontan a los orígenes de las trasmisiones y al que se le añaden unas 800.000 horas de emisión cada año dentro de la normativa de Depósito Legal.

A comienzos del siglo XXI, cuando el desarrollo de la web ha propiciado una nueva dimensión en el campo de las publicaciones y de las emisiones radiofónicas y de nuevas formas de reproducción, se aprovechan las oportunidades que la web ofrece para la distribución de recursos audiovisuales a través de la Web. La Ley de Depósito Legal se ha hecho también extensiva a los recursos en la web. Curiosamente los juristas franceses consideraron que era necesario que la responsabilidad fuera compartida entre la Biblioteca Nacional de Francia y el INA, que fue designado por la Ley como repositorio para albergar los recursos multimedia de la web y satisfacer la demanda de los recursos audiovisuales.

El INA comenzó a almacenar más de 8.000 emisiones radiofónicas referidas a Francia en la web en febrero de 2009. Esta colección en la actualidad complementa y continúa las colecciones de televisión e implementa nuevos recursos tecnológicos para su almacenamiento y se ha posibilitado e implantado nuevas formas de acceso a estos nuevos recursos en este nuevo contexto.

El INA La institución del patrimonio francés de contenidos sonoros y audiovisuales en el contexto de la web.

Desde su creación el INA ha tenido como misión la de recopilar, conservar y difundir las colecciones audiovisuales en Francia. En un principio estaba pensado para favorecer la conservación de los programas radiofónicos de acuerdos con las necesidades de los profesionales, pero pronto creció y se convirtió en el gran repositorio del patrimonio audiovisual francés cuando se suprimió el monopolio público.

En efecto, en Francia, en la segunda mitad de la década de los 80 se concedieron licencias de radiodifusión al recién nacido sector privado y ninguna norma aseguraba que su programación y emisión fuera recogida y guardada con fines patrimoniales. La comunidad académica presionó fuertemente y argumentó con firmeza que las emisiones de radio y televisión deberían conservarse para ser testimonio y constatación de los hechos acaecidos para generaciones futuras. Debido a su experiencia y legitimidad en la recopilación y conservación del patrimonio audiovisual, el INA fue designado por Ley, votada el 20 de junio de 1992[1], como responsable del Depósito Legal de la radio y la televisión. Unos veinte años más tarde, el Instituto está considerado como el mayor archivo mundial de radiodifusión y conserva más de 4 millones de horas de televisión y de radio que se remontan a las primeras emisiones con un incremento anual de unas 800.000 horas de programación de cien canales de televisión y 20 emisoras de radio con grabaciones digitales de las 24 horas de los siete días de la semana.

Es evidente que la era digital ha abierto nuevas perspectivas para las antiguas colecciones del INA y se inició un gran proyecto de digitalización en 1999 para salvaguardar y trasmitir en formato digital 830.000 horas conservadas en formato analógico en peligro de deterioro. En el 2015 todo este material sufrirá una migración para asegurar su conservación y su acceso a lo largo del tiempo.

Se negociaron aspectos referentes a temas de derechos de propiedad intelectual para parte de estas colecciones (30.000 horas) con el fin de poderlos poner en línea y facilitar su acceso al público en general, a los profesionales se ofrece un acceso restringido en línea de cualquier emisión de los que el INA es titular de los derechos de producción (alrededor de 1M de horas). Todo ello se puede buscar, ver y estudiar con fines de investigación en el centro (más de 4M de horas).

De forma paralela a la transferencia de soporte analógico al digital y a la conservación y trasmisión de imágenes la World Wide Web se convirtió en una herramienta esencial para publicar y acceder a diversos tipos de contenidos. Los organismos de radiodifusión y los nuevos agentes en el mundo del negocio de las telecomunicaciones (compresión de archivos digitales y acceso a banda ancha) para la distribución en línea de los archivos digitales pusieron en marcha un programa de plataformas cruzadas tanto en el acompañamiento y

[1] Ley de Depósito Legal para materiales audiovisuales votada el 20 de junio de 1992. http://legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000723108&categorieLien=id (último acceso 6 de mayo de 2012.)

en la puesta en marcha de una evolución- si no revolución -en el uso y en el consumo de recursos de vídeo desde múltiples dispositivos[2].

Dado el trepidante ritmo de las nuevas tecnologías en el mundo de la publicación, las normas relativas al Depósito Legal en Francia se hicieron extensivas a los recursos Web. Curiosamente, los legisladores franceses pensaron que era necesario compartir esta responsabilidad entre la Biblioteca Nacional de Francia BNF y el INA para garantizar la continuidad de sus respectivas colecciones[3]. El INA fue designado por tanto como el repositorio nacional de sitios WEB relacionados con los audiovisuales, de la forma más amplia posible, como los servicios bajo demanda de recursos audiovisuales desde las plataformas WEB. Una promulgación de la Ley se ha publicado recientemente para definir con precisión la misión conjunta y a la vez compartida de las dos instituciones[4].

La coherencia y la continuidad de las colecciones eran dos conceptos que suponían la columna vertebral jurídica tras el marco de la asignación al INA de esta nueva función. Sin embargo, su aplicación técnica y práctica partió, en parte, por su experiencia con las herramientas y prácticas llevadas a cabo para archivar el material de radiodifusión. Esta ponencia intentará proporcionar una visión general de los temas que entran en juego a la hora de seleccionar, adquirir, organizar, acceder, almacenar y conservar los archivos WEB que se enlazan, hacen referencia o se complementan con los archivos tradicionales de difusión.

Selección

En Francia, como en la mayor parte de los países, una agencia estatal, la CSA se encarga de regular las trasmisiones de la mayoría de las comunicaciones nacionales. Esto hace especial referencia a las trasmisiones de radio y televisión. Entre sus responsabilidades se incluyen la asignación y regulación de las distintas frecuencias necesarias para la trasmisión de las emisiones. Están por tanto sujetas a regulación y su número es limitado. Antes de que un nuevo canal salga al aire- algo que no ocurre todos los días-al contrario de lo que sucede en un sitio WEB-, se comunica públicamente antes de su emisión y entra dentro de los objetivos del INA el deber de archivarlo. El proceso generalmente se hace con anterioridad y transcurre sin incidencias. Pero para los sitios WEB se funciona de manera muy diferente. Éstos nacen y desaparecen sin previo aviso. La oficina de dominio WEB francés (AFNIC)

[2] La investigación en las diferentes formas de visualización ha demostrado que el vídeo en línea está sufriendo un crecimiento sin precedentes. Según la Internet Adverstising Bureau la radio tardó 38 años en llegar a 50 millones de usuarios , 13 años para los usuarios de TV, menos de 5 para Internet, menos de 2 para el vídeo por Internet…En el Reino Unido 27,3 millones de personas de los 38,5 millones de habitantes puso en línea su PC durante el año 2012 para ver el contenido en streaming (FuenteUKOM/Nielsen, Feb 2012)

[3] Ley de Derechos de Propiedad Intelectual en la Sociedad de la Información Actual, extensión de la Ley de Depósito Legal a la WEB votada el 1 de agosto de 2006http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350 (última consulta 6 de mayo de 2012)

[4] Promulgación de la ley referente a la legislación sobre Depósito Legal http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000025002022 (última consulta 6 de mayo de 2012)

puede de forma regular proporcionar una relación de aquellos sitios web de dominio fr., sin embargo éste sólo representa un 30% de la red francesa y no está clasificada atendiendo a sus actividades en la WEB.

Para seleccionar y evaluar los sitios WEB que se tienen que conservar hay que analizar la frecuencia y la profundidad de la misma junto a sus actualizaciones y tamaño y suponen el primer paso para aproximarse a un sitio WEB. Puesto que la WEB no tiene límites, es efímera y transitoria, definir la finalidad de la colección y su ámbito de actuación así como las fronteras de dominio que incluye, es un procedimiento esencial pero lento que consume mucho tiempo el desarrollar todo el proceso y que además implica que la aceptación se basa en una decisión humana. En base a los criterios objetivos establecidos en la Ley, los documentalistas del INA tienen que hacer un seguimiento de los sitios WEB relevantes y decidir cuales se tienen que conservar.

Siguiendo con los términos enunciados en la Ley la selección se basa tanto en la actividad del editor que se encuentra tras el dominio (actividad de radiodifusión) como también puede ser el hecho de que el dominio se refiera a la radio o la televisión (muchos blogs o sitios de fans de la WEB podemos encontrar en esta categoría) o que suministre vídeos bajo demanda o proporcionen accesos en línea o sin línea de programas (como una repetición, para ponerse al día, o contenidos originales de vídeo).

El proceso se inició hace más de dos años y comenzó en febrero recogiendo 3600 sitios WEB y se han incrementado hasta 10.000.
Cabe señalar que el ambicioso objetivo de la Ley de Depósito Legal de reunir todas las colecciones relativas a los dominios WEB hay que reconsiderarlo puesto que no es posible garantizar que todos los sitios WEB que se identifiquen y seleccionen se puedan archivar y más bien se trata de alcanzar el compromiso de hacer un mayor esfuerzo para tener un control mayor de las webs activas realizado por expertos profesionales. Se han experimentado métodos semiautomáticos de rastreo semántico para no dejar la decisión a un criterio humano aunque se ha demostrado que éstos no son lo suficientemente relevantes para cumplir con los objetivos propuestos por la INA. Más adelante demostraremos que ese supuesto de “mayor esfuerzo” también es necesario para adquirir y rastrear en la WEB.

Recopilar

Durante los años del período analógico el INA era el encargada de recopilar el material en un soporte físico siguiendo los métodos tradicionales de adquisición de cualquier biblioteca. Con el uso generalizado de las tecnologías digitales, ya durante 2001, se organizó una grabación a gran escala durante las 24 horas de emisión para más de 100 canales y el INA lo grababa desde los puestos donde se realizaba el trabajo y se almacenaba en su sistema de almacenamiento.

En cierto sentido las técnicas aplicadas y la infraestructura utilzada para la captura de contenidos WEB sigue, de alguna forma, el esquema de “estrategia invasiva” para la recogida de información aunque, a diferencia de la radio y la

televisión, la WEB es un medio no lineal que no “fluye” sino que, por el contrario, requiere de unas técnicas concretas y específicas.

La recopilación de contenidos directamente desde los servidores WEB es, sin duda, un enfoque general, mientras que el objetivo es el de simular todas las posibles actividades humanas dentro de las páginas WEB para generar el mayor número de respuestas y descargar todos los contenidos de una WEB remota. Estas técnicas se conocen comúnmente como “cosecha” o “rastreo” y las utilizan todos los motores de búsqueda para reunir y procesar los datos WEB. Las herramientas utilizadas son comúnmente denominadas “orugas” o “arañas”.
La palabra “araña” podría parecer un término arcaico u obsoleto, pero probablemente es el más apropiado para definir el método ya que alude a los múltiples caminos y encrucijadas que la herramienta automática tiene que descubrir, seguir o ignorar, con el fin de recopilar una parte del mundo WEB.

No se trata de una mera recopilación, sino que hablamos de una técnica muy precisa a partir de unas pautas por las que hay que seguir a través de los diferentes enlaces para descubrir y descargar el contenido de la WEB de acuerdo a los parámetros de rastreo. Todo esto sería relativamente sencillo si la WEB no estuviera en constante cambio, con nuevos enlaces impredecibles ya que unos aparecen mientras que otros, por el contrario, desaparecen con elementos nuevos o actualizados que se deben descubrir en cada nodo. De hecho, es posible predecir de alguna forma o tener conocimiento de cuando se producen estos cambios a través de los canales RSS actúan en calidad de alertas al igual que las vibraciones en la WEB la tela de araña la alerta de la existencia de un ser vivo para ser ingerido. Pero esto es más la excepción que la regla y la araña debe buscar continuamente su sustento a través de la WEB.

Esta analogía llega a sus extremos y, al igual que la araña debe girar y buscar y controlar sus propias redes mientras que los rastreadores que participan en el archivo de la WEB son los invasores, al igual que los parásitos que se expanden a través de una gigantesca WEB tejida por millones de arañas. Por otra parte cuando menos se espera surgen “trampas WEB” (intencionadas o no) que permiten que las herramientas caigan, como hormigas en un foso de arena. Pero ahora nosotros vamos a terminar con la similitud de los insectos y vamos a utilizar la palabra “rastreo”.

Puesto que el objetivo del INA para cumplir con el depósito legal está definido y nunca abarcará el gran número de sitios webs incluidos en las listas de los grandes rastreadores, el sistema de rastreo que se lleva a cabo es por niveles y se ha desarrollado para adaptarse lo mejor posible a la diversidad de los sitios WEB (referente a la periodicidad de la actualización, la profundidad de la información y sus características interactivas).

El sistema se basa en una arquitectura a dos niveles con un programador principal que manda órdenes a multitud de rastreadores, cada una debe ocuparse del sitio WEB asignado. Unos 500 o 1000 de estos rastreadores operan en una sola máquina.

El planificador, (la parte más superior de esta arquitectura de dos partes), evidentemente se ocupa de los aspectos referentes a la programación y a la configuración de cada sitio WEB. Utilizando una estrategia de muestreo simple supervisa la frecuencia de actualización y sus ritmos.
Se categorizan los sitios WEB (de una forma semiautomática dependiendo de las características de frecuencia (continuidad, actualizaciones por horas, días, semanas) y el rastreador se programa de acuerdo a las mismas.

Algunos sitios WEB tienen fuentes de distribución que a su vez enlazan con nuevos contenidos. Estas pautas se utilizan para comenzar rastreos específicos para una única página de un artículo, para su contenido, tan pronto como se publican dentro de un determinado sitio WEB, se hacen nuevas visitas de seguimiento automatizado de las actualizaciones y de los comentarios.

Este enfoque, trata de forma separada, por un lado, las estrategias de programación (el rastreo de la frecuencia y la profundidad de los parámetros coincidentes) y por otro lado con los problemas que puedan surgir (evitar las trampas de la WEB, rastreo de normas, la ejecución de las mismas).Esto permite el uso de diferentes rastreadores a la vez. Puesto que la WEB es una jungla perseguida por multitud de sistemas y técnicas heterogéneas, los rastreadores son propensos a errores y fallos mientras luchan para desarrollar una interacción.

Este enfoque específico sobre los multi rastreadores tiene como objetivo recoger varios tipos de contenido siguiendo la ya mencionada estrategia de “mayor esfuerzo” a la vez que mejora la calidad del archivo. Como al menos tres rastreadores se pueden conectar a un planificador se puede dedicar cada uno de ello a una tarea específica:

Phagosite: es un rastreador para fines generales que puede manejar webs muy grande y no requiere de grandes recursos informáticos.

Fantomas : es un rastreador más específicos basados en el kit Phantom-JS Web que utiliza las mismas funciones básicas de los navegadores Google Chrome o Apple Safari. Este rastreador es capaz de rastrear las páginas WEB “2.0” con complicadas estructuras de Java Script sin ser demasiado estricto con los recursos del ordenador.

Crocket: Se basa en el navegador Firefox y es capaz de rastrear sitios WEB complejos y ricos (ricos recursos y ricas interacciones) pero sin embargo es potentes para los recursos del ordenador.

Además, como el vídeo es una parte importante (tanto por el número como por el tamaño) del archivo, se han desarrollado una serie de rastreadores para descargas vídeos UGC de YouTube o Dailymotion, y recoger vídeos disponibles en streaming.

Este sistema exclusivo de rastreo personalizado ha estado funcionando continuamente desde 2009 a un ritmo de 6 mil millones de peticiones al año.

Las herramientas se actualizan y mejoran continuamente para adaptarlas al mundo siempre cambiante y en proceso continuo de maduración de Internet.

Descripción, metadatos y acceso

Los archivos fílmicos y sonoros tanto digitales como los de la WEB requieren de unos requisitos tecnológicos para acceder a sus contenidos que no existían con anterioridad a los contenidos tradicionalmente publicados. El contenido de un libro puede estar disponible inmediatamente, siempre se puede leer, mientras que un soporte físico como una película, una cinta, un disco o los archivos digitales de información necesitan ser configurados, calculados e interpretados para hacerlos comprensibles a los humanos.

En el INA las tareas de los documentalistas se han reorientando teniendo en cuenta la gran cantidad de datos que estos soportes conllevan tanto para el desarrollo de la emisión como para los contenidos de Internet y la extracción y gestión de metadatos se han convertido en tareas prioritarias.

Sin embargo, siguiendo con un enfoque tradicional para organizar las colecciones los sitios WEB se documentan y catalogan según las taxonomías específicas que tratan de unir las emisiones con las colecciones de la WEB.

Los contenidos de los archivos WEB se indexan automáticamente para permitir un acceso aleatorio partiendo de una URL y una fecha. El almacenamiento en disco, los ficheros y los metadatos conforman el archivo. Debido a las características de un archivo digital que almacena una información diferenciada y, debido a la naturaleza de la WEB como una herramienta de publicación el acceso a las páginas WEB “archivadas” (una reciente estimación considera que existe un promedio de unos 50 archivos individuales que componen una página), implica la reconstrucción de un proceso de publicación a través del acceso a los archivos que se han rastreado en un determinado momento y que permiten su visualización dentro de una página reconstruida.

La herramienta de búsqueda funciona bajo el buscador Firefox. Desde una determinada captura es posible proseguir o retroceder en el tiempo y, finalmente, mostrar todas las versiones disponibles. La mayoría de las interacciones permanecen disponibles. La mayoría de las relaciones permanecen activas (enlaces de navegación y relaciones básicas que todavía representan el 95% de la experiencia del usuario), algunas no se establecen por razones técnicas (contenidos interactivos de Flash o interacciones complejas de Javas Script incluyendo algunos reproductores de vídeo. Evidentemente, el objetivo es capturar todas y mostrar el denominado “look and feel “ o el contenido de las páginas tal y como ellas se concibieron, sin embargo este intento no ha conseguido todavía un éxito absoluto, la mayoría de las veces debido a problemas técnicos.

Por ejemplo, algunas las relaciones que se han perdido necesitan ser recreadas para permitir una experiencia completa de su navegación por ejemplo:

Vídeos que no se pueden reproducir en el dispositivo creado para su reproducción se pueden reproducir con un dispositivo externo y se pueden proporcionar otras posibilidades adicionales como la búsqueda dentro del vídeo.

Como las búsquedas se realizar a través de Google o Bing no se pueden almacenar (una simulación de todas las posibles simulaciones es algo imposible). Se ha desarrollado un determinado motor de búsqueda para buscar en el archivo WEB del INA que se ha implementado para controlar la dimensión del tiempo en el archivo y agrupar los resultados duplicados o los similares a los duplicados.

Dowser es nuestro motor mágico de búsqueda que permite al usuario el acceso a cualquiera de los contenidos archivados desde las preguntas realizada según "la vía Google". Se utilizan historiogramas para ayudar al investigador a navegar a través de fechas y datos a través de cualquier consulta realizada.

La autenticidad mantiene la legitimidad de un archivo pero esta idea ha sido muy criticada en un entorno digital. La idea pura de un documento "original" ha desaparecido pronto y los datos digitales surgen para ser copiados, migrados o manipulados. El acceso a los contenidos WEB archivados pueden ofrecer continuas presentaciones y visionados desde sitios WEB pero no proporcionan una versión pura de una página WEB puesto que ésta no existe. La opción del INA es la de informar al usuario de cualquier posible modificación del contenido original tanto en línea como en un contexto (una discontinuidad temporal entre la página a la que se ha hecho el enlace y su referente, el vídeo reproducido en un reproductor externo al lugar donde se albergaba originariamente, los contenidos que no aparecen debido a los límites del rastreador etc), haciendo hincapié en el hecho que los documentos WEB archivados no son en realidad un original sino una copia incompleta del original . De nuevo: "hacer lo mejor posible" es el camino correcto a seguir.

Almacenamiento y preservación

El contenido que se publica en la WEB es mayoritariamente de formato digital y no existe en ningún otro lugar. Puesto que contendrá muchos de los registros

históricos de nuestro tiempo y será un testimonio de nuestra sociedad actual la preservación a largo plazo es un tema crucial así como el almacenamiento de archivos de gran tamaño que permitan la migración de una tecnología a otra como es la de las cintas magnéticas al almacenamiento en disco. Como todo el mundo que trabaja en la WEB sabe, el archivo necesita desarrollar una estructura mejor para la preservación de contenidos WEB a largo plazo y esta estructura está todavía en fase de desarrollo por lo que la mayoría de las instituciones centraron sus esfuerzos en proporcionar el mejor acceso posible de sus colecciones para continuar con el proceso de recopilación. Vamos a continuación a exponer los principales temas que se debaten dentro de la comunidad de los archiveros WEB y que todos los especialistas que trabajan con archivos digitales.

Asumiendo que la integridad de los datos y el nivel de bits se mantiene ¿los datos almacenados se podrán reinterpretar de forma correcta en un futuro? Tal y como sucede en la lectura de una película o una cinta en la que se deben mantener, en el caso de la migración de formatos, los archivos es igualmente esencial el mantenimiento de los buscadores y de los plugins.

Conservación a largo plazo (usando migraciones a corto plazo)
La preservación de los archivos a largo plazo es similar a lo que sucede con el trabajo de los archivos digitales utilizando las migraciones de datos a corto y medio plazo (aproximadamente utilizando estrategias para un plazo de 20 años).
En el INA se conservan dos copias de los archivos almacenados en diferentes generaciones de discos para el almacenamiento de datos y dos copias del backup se guardan en cintas fuera de línea.
Esto se basa en la migración de los archivos almacenados en un nuevo soporte de disco para el almacenamiento que se hace cada 4-5 años y en una cinta que también se hace cada 4-5 años.

Almacenamiento en disco
La cuestión referente a los requisitos de mantenimiento es mucho más importante que desde las páginas de un texto formateado en (HTML) en kilobytes. Por el contrario la necesidad de almacenar una cantidad creciente de audio y vídeo necesita de muchos megbytes. A diferencia del almacenamiento en cinta el almacenamiento en disco conlleva unos costes más elevados asociados a la potencia y a la refrigeración. Las tasas de fiabilidad y de fracaso de la tecnología de almacenamiento en disco también son factores que han de tenerse en cuenta y que limitan el uso de estos dispositivos a 3 o 5 años. Para facilitar la migración y ahorrar en el coste se dobla la capacidad de almacenamiento en cada migración (por ejemplo discos de 1 a 3,5 TB).

LTO (Cinta lineal abierta)
LTO es una cinta magnética de tecnología para el almacenamiento de datos tecnológicos desarrollada a finales de 1990 en virtud de una iniciativa de estándares abiertos en contraposición a los formatos de cintas magnéticas patentados que se usaban en esos años. A partir del 2000 una norma define un cartucho LTO como un cartucho de tamaño estándar que incrementa su capacidad de almacenamiento en más de 8 generaciones, es decir por un

periodo de unos 20 años (2000-2017). La capacidad es aproximadamente el doble en cada nueva generación, facilitando las estrategias de migración y permitiendo la normalización de almacenamiento en los depósitos.
A pesar de que las cintas pueden almacenarse teóricamente por más de 25 años no se tiene la seguridad que los proveedores de este tipo de cintas vayan a producir cintas que tengan la capacidad de leer las ahora existentes. De hecho la especificación LTO solo exige como norma la compatibilidad es decir que cada nueva versión sea compatible con las dos versiones anteriores (por ejemplo cintas LTO4 puedan leer las LTO 2).
Por lo tanto para migrar de forma fiable los datos hay que migrarlos al nuevo formato cuando sea completamente compatible y económicamente viable. Por ejemplo a pesar de que las unidades LTO5 están disponibles en Q2 2012 es el precio el que no favorece la viabilidad.
De forma similar hay que ver la viabilidad económica cuando se quiera migrar para de esta forma maximizar la eficacia y el tiempo de duración. Lo mejor es hacerlo entre dos generaciones de LTO es decir, a través de cuadruplicar la capacidad de cada cartucho (por ejemplo migrar 4 cartuchos de LTO4 a 1 de LTO6)

Verificación
Después de la creación de un archivo histórico hay que verificar su contenido a través de diversas herramientas. Si se considera que el contenido del archivo es válido se hace una verificación del archivo que se almacena por separado en una base de datos aparte. Las verificaciones (SHA) también se almacenan con sus archivos asociados cuando se almacenan en soportes magnéticos. De esta forma el contenido y la validez de los archivos se puede comprobar periódicamente y también si existe una copia dañada se puede sustituir a partir de un backup.

Archivos de larga duración ¿emulación o migración?
Es necesario considerar los archivos WEB desde dos puntos de vista:

- El “Look and Feel” de las páginas WEB es la forma “estética” en la que se trasmite el estilo, la comunicación y el instinto de su creador.
- Los datos que componen sus elementos. Las palabras escritas que expresan las ideas del autor, las imágenes, los sonidos y los vídeos.

La primera parece estar más abierta a las estrategias de tipo de la emulación. En la actualidad un numeroso grupo de organizaciones e instituciones están ejecutando proyectos para clasificar y emular a los navegadores y plugins antiguos.
Una forma de verificar que la emulación está funcionando correctamente es la de probar un vídeo o cualquier instantánea de la WEB.

La segunda fórmula se presta más a una estrategia de tipo de migración donde en el punto de la captura (o incluso dentro de unos años) se puede identificar un archivo estándar a medio plazo (unos 20 años) después de la migración de los datos.

En el ejemplo anterior PDF o ASCII se pueden elegir para cualquier forma de texto, JPEG2000 para cualquier tipo de imágenes, AIFF para el sonido H264

para cualquier formato de vídeo. Estos son en la actualidad los formatos estandarizados que se usan en el INA para la migración largo plazo de los formatos de audio y vídeo.
Este tipo de enfoque también podría utilizarse para los metadatos que se han extraído donde una serie de herramientas de software se pueden ejecutar para ayudar a identificar los archivos y los formatos a la vez (por ejemplo (DROID, JHOVE o Apache Tika) y almacenar los resultados en formatos estandarizados y organizados.
En última instancia el objetivo es el de capturar la mayor cantidad de datos posibles de una página WEB en el momento de su captura. Incluso con la mejor migración/o estrategias de emulación no hay garantía de que una página capturada sea posible recuperarla en la forma en la que fue almacenada originariamente. Los aspectos relativos a la visualización de la página se pueden restaurar pero lo que es posible es la navegación no se puede realizar de la misma manera.

Sin embargo, combinando ambas estrategias, se permitirá la utilización de una serie de herramientas para el usuario futuro.
Por ejemplo él o ella podrá ser capaz de ver una página creada por un emulador que no se puede leer asociada a un documento en una página reconstruida. Sin embargo una emulación se puede comparar con una imagen de la página con la finalidad de dar una idea de su forma visual primitiva y, a continuación, en referencia a los datos extraídos (por la migración) recuperar también el contenido original.

Conclusión

No tenemos los medios para calcular realmente el impacto a largo plazo del almacenamiento digital de todo el conocimiento pero las academias, las instituciones nacionales y las academias están tratando de hacer todo lo posible para garantizar que se mantenga vivo por mucho tiempo. El archivo WEB es un paso hacia delante que permite el almacenamiento de una información que por su propia naturaleza tiene una vida relativamente corta vivo en una página WEB, manteniéndolo al día y utilizando para ello numerosos formatos, manteniendo la interactividad de en el archivo o activar “la experiencia del usuario” tal y como estaba previsto en su origen. Ahora es una actividad compartida y aunque los enfoques y las opciones pueden ser diferentes de una institución a otra, la cooperación entre ellos y el fomento del diálogo IIPC de intercambio de experiencias, la participación de los usuarios, académicos y profesionales de muy diferentes países para que la historia de la World Wide Web y su contenido pueda ser escrito por generaciones futuras.