Gestión y preservación de documentos de archivo en ambientes web

7

Traducción al español:
Alicia Barnard, Alejandro Delgado
y Juan Voutssás

Cuadernos Digitales de Archivística

Serie: Temas fundamentales de preservación digital

ARCHIVO GENERAL DE LA NACIÓN

Dirección General: Mercedes de Vega
Dirección General Adjunta de Administración: Alba Alicia Mora Castellanos
Dirección de Publicaciones y Difusión: María Fernanda Treviño Campero
Departamento de Publicaciones: Esther Pérez Guzmán

Coordinación Editorial: Esther Pérez Guzmán
Diseño y formación: Alejandro Amaro Rosas
Diseño de portada: Alejandro Amaro Rosas
Corrección de estilo: Ma. del Carmen Gutiérrez Haces y Francisco J. González Ruiz
Asistencia editorial: Roberto del Vecchyo Calcáneo

Gestión y preservación de documentos de archivo en ambientes web
Traducción al español: Alicia Barnard, Alejandro Delgado y Juan Voutssás

Archivo General de la Nación
Eduardo Molina núm. 113
Col. Penitenciaría
Deleg. Venustiano Carranza, C.P. 15350
Ciudad de México

Primera edición: octubre de 2016

Contenido

Agradecimientos

Muchas personas contribuyeron a la creación de los ocho módulos que integran esta serie, en particular, los estudiantes de doctorado de la Universidad de la Columbia Británica, Elizabeth Shaffer, Corinne Rogers, Donald Force y Elaine Goh, quienes elaboraron los borradores de los contenidos basados en los trabajos de InterPARES 1 y 2, así como los casos de estudio desarrollados en InterPARES 3.

También agradecemos a los numerosos asistentes de investigación quienes elaboraron casos de estudio para todos los módulos así como al equipo de InterPARES en Canadá, a un sinnúmero de investigadores internacionales involucrados con este proyecto y, por supuesto, a su directora, Luciana Duranti.

Finalmente, nuestra gratitud a todos aquellos que revisaron y comentaron los módulos, con una mención especial a los investigadores John McDonald, consultor de administración de información (módulos 1, 2, 7 y 8), Jim Suderman, director del despacho de acceso a la información de Toronto (módulo 3), Evelyn McLellan, archivista de sistemas de Artefactual Systems, Inc. y Paul Hebbard, archivista administrador de documentos de archivo de la Universidad Simon Fraser (módulo 6).

Prefacio a la edición en inglés

Digital Records Pathways: Topics in Digital Preservation es una iniciativa educativa desarrollada en conjunto por el International Congress on Archives (ICA) y The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) con el propósito de ofrecer capacitación a archivistas y profesionales que manejan documentos en cuanto a la producción, la administración y la preservación de documentos de archivo digitales auténticos, fiables y usables. El programa asume que el lector cuenta con una sólida base en cuanto a los conceptos fundamentales de la administración de archivos y en la teoría archivística, y sobre esa suposición se elaboró esta serie modular.

La serie está formada por ocho módulos más un glosario en donde se ha conjuntado terminología de acuerdo con la base de datos del ICA. Ésta aborda los conocimientos teóricos y prácticos necesarios para establecer el marco de referencia, la estructura de gobernanza y los sistemas requeridos para administrar y preservar documentos de archivo digitales, a través de su ciclo de vida. Cada módulo se refiere, específicamente, a un tema relevante para la administración o la preservación de los documentos de archivo. Todos los módulos se han diseñado de tal manera que pueden ser estudiados en forma independiente o en conjunto.

Prefacio a la edición en español

Desarrollar materiales educativos con el fin de apoyar las tareas de preservación digital en instituciones y organizaciones fue uno de los objetivos del Proyecto de Investigación Internacional para la Preservación de Documentos de Archivo Electrónicos, InterPARES 3 (2007-2012), el cual dio como resultado la serie en inglés de ocho módulos de capacitación con el título Digital Records Pathways: Topics in Digital Preservation.

No obstante que los archivos digitales –también llamados electrónicos–, se producen en la actualidad en volúmenes insospechables, el conocimiento de los archivistas y gestores de documentos en cuanto a la producción, conservación y preservación de los mismos aún es limitado, entre otros factores, a causa de la escasez de materiales de capacitación en idioma español.

Lo anterior fue el motivo para que Alicia Barnard y Juan Voutssás, miembros del Team México que formó parte del Proyecto InterPARES 3, junto con Alejandro Delgado, de España, se dieran a la tarea de traducir a nuestro idioma los ocho módulos de la citada serie, los cuales fueron publicados inicialmente en formato electrónico por el Proyecto InterPARES 3.

El Archivo General de la Nación se une a este esfuerzo para lograr una mayor difusión de temas y tópicos sobre la preservación de archivos digitales en el entorno de los archivos de nuestro país y de aquellos de habla española en Latinoamérica, y presenta una nueva versión electrónica en español de los módulos de la mencionada serie, en espera de que coadyuven a la mejor comprensión y entendimiento de la preservación de archivos digitales y el ambiente donde los mismos se producen, conservan y preservan.

Mercedes de Vega

Acerca de ICA e InterPARES

The International Council on Archives (ICA) y The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) tienen el compromiso de crear materiales didácticos para la educación continua de archivistas y administradores de documentos de archivo, construir conocimiento básico, diseminar los nuevos hallazgos y dotar a los archivistas y profesionales de los documentos de archivo del conocimiento y las competencias necesarias especializadas para la administración y la preservación de documentos de archivo digitales.

El ICA (www.ica.org) está dedicado al manejo eficaz y a la preservación de documentos de archivo, así como al cuidado y uso del patrimonio archivístico mundial y su representación, por medio de profesionales en todo el planeta. Los archivos son un recurso increíble: son un subproducto documental del quehacer humano y, por tanto, testigos irremplazables de eventos pasados, puntales de la democracia, de la identidad de individuos y comunidades, así como de los derechos humanos; pero también son frágiles y vulnerables. El ICA se esfuerza por proteger los archivos y asegurar su acceso por medio de la asesoría, el establecimiento de estándares, el desarrollo profesional y el impulso del diálogo entre archivistas, líderes, productores y usuarios de archivos.

El ICA es una organización neutral, no gubernamental; sus miembros operan por medio de las actividades propias de cada membresía. Por más de sesenta años, el ICA ha unido a instituciones archivísticas y practicantes, a lo largo del mundo, para asesorar acerca de la buena administración archivística y la protección física del patrimonio registrado, para producir estándares reconocidos, buenas prácticas e impulsar el diálogo, el intercambio y la diseminación del conocimiento y experiencia más allá de fronteras internacionales. Con aproximadamente

mil quinientos miembros en 195 países y territorios, el credo del Consejo ha sido aprovechar la diversidad cultural de sus integrantes para entregar soluciones eficaces y una profesión flexible e imaginativa.

El proyecto InterPARES (www.Interpares.org), pretende desarrollar conocimiento original y esencial para la conservación, a largo plazo, de documentos de archivo producidos y almacenados en formatos digitales, así como proveer una base sólida para estándares, políticas, estrategias y planes de acción capaces de asegurar la longevidad de los materiales documentales y la capacidad de sus usuarios para confiar en su autenticidad. InterPARES se ha desarrollado en tres etapas:

- InterPARES 1 (1999-2001). Esta etapa se enfocó en el desarrollo de la teoría y los métodos que pudiesen asegurar la preservación de la autenticidad de los documentos de archivo producidos y conservados en bases de datos y sistemas de gestión de documentos de archivo, durante el curso de las actividades propias de su administración. Los hallazgos de esta etapa presentaron el punto de vista del preservador de los documentos de archivo.
- InterPARES 2 (2002-2007). Se continuó investigando acerca de temas relativos a la autenticidad, fiabilidad y exactitud durante todo el ciclo de vida de los documentos de archivo, desde su producción hasta su conservación permanente. Se enfocó en aquellos documentos de archivo producidos en entornos digitales dinámicos e interactivos a lo largo de actividades artísticas, científicas y gubernamentales.
- InterPARES 3 (2007-2012). Se construyó sobre la base de los hallazgos de las primeras dos etapas en conjunto con otros proyectos de preservación digital de distintas partes del mundo. Se llevó la teoría a la

práctica al trabajar con archivos y unidades archivísticas dentro de organizaciones que tuvieran recursos humanos y financieros limitados, con el fin de implementar en ellas programas sólidos de gestión y preservación de archivos.

Público objetivo

El público objetivo para el cual está destinado este programa se compone de archivistas, gestores documentales y profesionales de la gestión archivística, interesados en ampliar sus capacidades en la administración de documentos de archivo digitales. En conjunto, los módulos conforman todo un paquete de recursos documentales para la educación continua de profesionales, con especial énfasis en aquellos temas que impactan en la preservación de documentos de archivo, auténticos, fiables y exactos.

Cómo usar la serie

Cada módulo de la serie está formado por conocimiento teórico y metodológico, así como por aplicaciones prácticas ilustradas en casos de estudio y escenarios modelo. Si bien los módulos fueron desarrollados por el equipo canadiense de InterPARES y, por tanto, ejemplificados en un contexto propio a aquél, son adaptables a un dominio específico o ámbito jurídico. Para una mayor aplicabilidad se han traducido a los idiomas de los miembros del ICA.

Los módulos pueden estudiarse por separado o en conjunto, de acuerdo con cada necesidad o interés, pues abarcan un rango amplio de competencias requeridas; pueden ser estudiados individualmente u ofrecerse a grupos como asociaciones profesionales o instituciones de capacitación laboral. Algunos de los módulos incluyen formularios

que pueden adaptarse a universidades o asociaciones profesionales para el desarrollo de cursos curriculares, o como materiales de capacitación para estudiantes y profesionales de la gestión o preservación documental digital. Las universidades y asociaciones profesionales son libres de adaptar los materiales para desarrollar sus propios cursos curriculares o de capacitación. Se sugieren recursos adicionales en la web que se identifican a lo largo de los módulos con el icono:

o bien, cuando se trata de información complementaria ubicada en anexos o en los mismos módulos de la serie, se distinguen con la figura:

Objetivos

Los módulos tienen los siguientes objetivos:

- Aportar recursos educativos basados en investigación actual sobre temas de administración de archivos digitales para beneficio de miembros de asociaciones profesionales relacionadas con esa temática.
- Proporcionar a los profesionales de archivos, con conocimiento teórico y procedimental, habilidades estratégicas necesarias para desarrollar, implementar y supervisar un sistema de gestión o de preservación documental.
- Ilustrar conceptos teóricos con aplicaciones prácticas mediante ejemplos reales extraídos de casos de estudio, asociados con contextos administrativos y tecnológicos específicos.

- Proporcionar contenidos y estructura a programas educativos universitarios para implementar cursos sobre administración o preservación de archivos.

Arquitectura modular del programa

Los primeros dos módulos presentan los fundamentos de todo programa de preservación de documentos de archivo digitales; proporcionan los conocimientos propedéuticos sobre los demás módulos. Los siguientes tres módulos tratan temas generales contemporáneos que competen a la preservación digital: el papel de la cultura organizacional, una visión general de los metadatos y de la valoración en el contexto de la administración de documentos de archivo fuera del sistema de gestión documental Electronic Recordkeeping Management System (ERMS). En los tres últimos módulos se abordan temas específicos de interés contemporáneo: la administración de correos electrónicos, la preservación de documentos de archivo en ambientes web, y los temas emergentes acerca del creciente auge del cómputo en la nube (tabla 1).

Tabla 1 Arquitectura modular del programa

Módulo	Aspecto
1. Un marco de referencia para la preservación digital. 2. Desarrollo de políticas y procedimientos para la preservación digital.	Fundamentos
3. Cultura organizacional. 4. Resumen de metadatos. 5. Estrategias de valoración.	Generalidades
6. Correo electrónico. 7. Sitios web. 8. Cómputo en la nube.	Específico
Base de datos internacional, terminología.	Fundamentos

Cada módulo contiene todos o algunos de los siguientes elementos:

- Panorama del tema y alcance del módulo.
- Objetivos y aprendizajes esperados del módulo.
- Metodología o procedimientos para la aplicación y desarrollo del módulo.
- Formularios (cuando apliquen) para facilitar la implementación del módulo.
- Ejemplos, casos de estudio o escenarios (cuando apliquen) con situaciones reales acerca del tema.[1]
- Ejercicios de los puntos clave del aprendizaje.
- Preguntas de revisión que optimicen la comprensión y entendimiento del tema.
- Recursos adicionales.
- Lecturas, estándares y otros recursos de referencia.

Cuando se ha considerado apropiado, se hace la distinción de la administración y preservación de documentos de archivo activos en contraste con las responsabilidades relativas a éstos que ya no son requeridos para actividades cotidianas de la organización y que serán preservados por su productor o por un tercero de confianza.

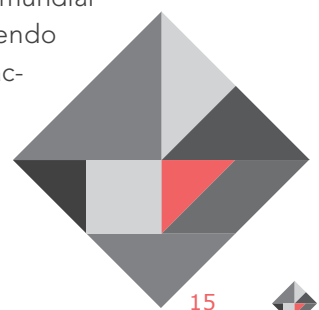
[1] Los ejemplos y casos de estudio citados en los módulos provienen de casos reales de InterPARES 3 y tienen como propósito apoyar la experiencia de aprendizaje del módulo. Si bien reflejan los hallazgos de investigación del proyecto, no necesariamente deben ser tomados como plantillas para ser aplicadas a pie juntillas en todos los casos. Cada organización (productor o preservador) es diferente y la preservación de sus documentos de archivo debe tomar las mejores prácticas desde una perspectiva práctica en cuanto a la viabilidad de una cierta implementación.

Alcance

La serie comprende los siguientes ocho módulos:

Módulo 1 Un marco de referencia para la preservación digital.
Módulo 2 Desarrollo de políticas y procedimientos para la preservación digital.
Módulo 3 Cultura organizacional y sus efectos en la administración de archivos.
Módulo 4 Breviario de metadatos.
Módulo 5 Estrategias para lograr el control de los documentos de archivo digitales en ambientes de red distribuidos.
Módulo 6 Administración y preservación de correo electrónico.
Módulo 7 Administración y preservación de documentos de archivo en ambientes web.
Módulo 8 Introducción al cómputo en la nube.

Para asegurar un entendimiento generalizado y reducir un potencial riesgo de confusión que pudiese surgir de prácticas regionales o jurisdiccionales, estos módulos están apoyados por una base de datos de administración de archivos la cual refleja los usos habituales y prácticos en 16 idiomas. Esta base, desarrollada conjuntamente por el ICA e InterPARES está disponible en www.web-denizen.com/. Dicho recurso dinámico continuará creciendo y desarrollándose en la medida en que los miembros de la comunidad archivística mundial puedan participar agregando o enriqueciendo las definiciones usadas en su región de práctica. Pueden verse ciertos términos específicos, en breves glosarios existentes en cada módulo, aún no incluidos en la base de datos.



Introducción

La naturaleza ubicua de la web ha dado como resultado la proliferación de sitios, muchos de los cuales contienen información y documentos de archivo valiosos para las organizaciones. Los sitios web han pasado de ser inicialmente repositorios de información a convertirse en sitios donde se producen documentos de archivo, tienen lugar transacciones y trámites, y se publica la información, lo cual plantea cuestiones únicas sobre la gestión documental y la preservación para este ambiente. Debido a la naturaleza de la web, puede ser difícil asegurar la autenticidad de los documentos de archivo que se producen y residen en sus sitios.

Muchas de las grandes organizaciones han sido fundamentales en el desarrollo de métodos de captura, herramientas y métodos de preservación.[2] Algunas de estas herramientas y técnicas son fácilmente adaptables a las necesidades de las organizaciones o los programas archivísticos pequeños o medianos. El Archivo Internet se ha desarrollado como una de las soluciones de código abierto para operaciones de recopilación o recolección que no requieren un costo monetario, pero sí un conocimiento tecnológico suficientemente extenso. Los Archivos Nacionales de Reino Unido han llevado a cabo investigaciones sobre los mejores medios de almacenamiento, una guía simple para archivar sitios web, así como sobre los formatos óptimos para la producción de datos. Los Archivos Nacionales de Australia han investigado sobre los requisitos de metadatos clave para manejar, de forma efectiva, todos los documentos de archivo digitales, incluyendo aquellos basados en web. También han investigado sobre soluciones para archivar (integrar) evidencia de documentos de archivo basados en web en sitios con cambios frecuentes cuando se tienen buscadores inusuales.

[2] Entre las organizaciones más grandes que actualmente preservan sitios web están la Biblioteca del Congreso de Estados Unidos de Norteamérica, el Archivo Internet, los Archivos Nacionales de Reino Unido y los Archivos Nacionales de Australia.

Objetivos y metas de este módulo

El objetivo de este módulo es introducir cuestiones clave involucradas en la gestión y preservación de sitios web, incluyendo la identificación de documentos de archivo en ambientes web. Esto ayudará a los usuarios a reconocer las necesidades de gestión y preservación de estos documentos de archivo, identificar las cuestiones involucradas en el manejo de documentos de archivo en un ambiente web e introducir estrategias para la preservación de estos sitios en este ambiente.

Aprendizajes esperados

Al completar este módulo el estudiante será capaz de:

- Identificar documentos de archivo en ambientes web.
- Comprender cuestiones clave involucradas en la gestión y preservación de documentos de archivo en ambientes web.
- Entender las diferentes estrategias para la preservación de sitios web.
- Saber dónde localizar información adicional y recursos que facilitarán cómo manejar y preservar documentos de archivo en ambientes web.

Terminología

Los conceptos clave que se emplean en este módulo son:

Compatibilidad retrospectiva: compatible con modelos o versiones anteriores del mismo producto. Se dice que una nueva versión de un programa

tiene compatibilidad con versiones anteriores si el mismo puede utilizar archivos y datos producidos con versiones más antiguas del mismo programa. Se argumenta que una computadora es compatible si puede correr el mismo programa o aplicación de manera semejante que la versión anterior de la computadora.

Métodos de colección del cliente: fuente desde la cual el sitio web es colectado para su preservación. Las recopilaciones del cliente son recopiladas mediante un navegador.

Transferencia directa: adquisición de una copia de los datos directamente desde la fuente original. Requiere acceso al servidor principal. Involucra copiar datos desde el servidor y transferirlos a la institución responsable de su recopilación.[3]

Recolección remota: técnica más común para archivar un sitio web, utiliza los buscadores de este entorno para automatizar el proceso de acumulación de páginas web. Los rastreadores normalmente "ven" las páginas web de la misma manera que los usuarios con su navegador y, por tanto, proporcionan un método simple comparativo para recolectar contenidos de forma remota.

Métodos de recolección desde el servidor: fuente desde la cual el sitio web es recolectado para su preservación. Las acumulaciones en un servidor son reunidas a través del servidor web.[4]

Rastreador web: programa de computadora que busca en la web de forma metódica y automática.

Sitio web espejo: copia exacta de un conjunto de datos. Esencialmente trabaja como una copia impresa de un sitio web.

[3] Brown, Adrian, *Archiving Websites.*
[4] *Idem.*

El proceso para producirla origina una copia del original pero no captura los datos asociados.

Véase la base de datos internacional de terminología del Consejo Internacional de Archivos (ICA por sus siglas en inglés) en http://www.Web-denizen.com para más términos relacionados con el tema de este módulo.

Estrategia de preservación de sitios web

No existe una solución única definitiva para ser aplicada a la preservación de sitios web. Las estrategias dependerán de una variedad de factores que incluyen la presencia (o ausencia) de documentos de archivo en el sitio, la propiedad del contenido, las capacidades tecnológicas, los costos y las habilidades de almacenamiento. Por tanto, existen varios planes de acción que pueden concebirse para la preservación a largo plazo de un sitio web institucional. Los planes de acción abarcan desde soluciones relativamente baratas que simplemente preservan instantáneas de sitios web hasta soluciones sumamente técnicas, promovidas por ciertas empresas de bases de datos, altamente efectivas porque abordan el dinamismo de sitios web subyacentes.

Hay herramientas disponibles para archivar sitios web. Aquella que se seleccione dependerá de qué tanta información desea preservar la organización, las habilidades técnicas del personal y un meticuloso análisis de riesgo.

Existen muchas consideraciones para una organización que emprende un programa de preservación de sitios web. Los factores incluyen: habilidades tecnológicas, administración de derechos, capacitación, descripción de recursos, documentación y acceso; selección de formatos, verificaciones de validación, planes de recuperación en casos de desastre, medios de almacenamiento, estándares y método de captura de sitios web.

Capacidades tecnológicas

Algunas estrategias requieren un conocimiento intensivo del ambiente tecnológico a fin de que pueda ser implementado, mientras otras necesitan una cantidad mínima

de conocimiento para implantación y logro de éxito. Es relativamente simple tratar los sitios web que integran documentos estáticos e incorporan muy poca o ninguna interoperabilidad. Sin embargo, el tratamiento es complejo para aquellos sitios que incorporan altos niveles de interactividad y comprenden páginas generadas dinámicamente y se han probado que resultan difíciles de archivar de manera efectiva.

Es importante que las personas responsables de la preservación de materiales digitales cuenten con conocimiento de lo que involucra el proceso. El responsable debe contar con suficientes nociones para tener comunicación informada con quienes están involucrados en la estrategia de preservación; también debe ser capaz de establecer requisitos realistas a un tercero. Además, dentro de la organización, se deberán considerar las siguientes características cuando se evalúen las estrategias apropiadas de preservación. Entre ellas podemos distinguir:

- Tipo de sitio web (por ejemplo, estático o dinámico).
- Ubicaciones en espacio del servidor y disponibilidad.
- Capacidades de respaldo.
- Sistema de cómputo en uso.

Política y requisitos de gestión archivística

Como profesional de documentos de archivo en su organización, usted debe estar al tanto de los principios y prácticas de administración de documentos de archivo que se practican. Este módulo, y los otros de la serie, ofrecen la oportunidad para revisar la efectividad de estas prácticas y la vigencia de cualquier documentación. Por ejemplo, se desea revisar los requisitos de concentración y disposición para ciertas series documentales; actualizar políticas y procedimientos relacionados

con las funciones de administración o reconsiderar las fortalezas y debilidades de las herramientas educacionales, las presentaciones o documentos que usted usa para capacitar a los empleados acerca de la importancia y prácticas de la administración de documentos de archivo.

Las políticas, procedimientos y criterios para un programa de preservación de sitios web son críticos en el ambiente digital emergente. Éstos aseguran que los propósitos y objetivos de la institución son considerados y revisados cuidadosamente, que el desarrollo de colecciones sustenta la misión y prioridades institucionales y aseguran la rendición de cuentas de los organismos que financian y la comunidad en general. Los elementos a considerar en una política son: declaración, metas y objetivos, documentos relacionados y legislación, alcance, personas responsables de la implementación, alcance de las compilaciones, cobertura y una descripción de los tipos de recurso digital aceptados, criterios de rechazo, criterios de evaluación, viabilidad y niveles de la compilación. Todos éstos pueden ser divididos en más de una política.

Véase el módulo 2, Desarrollo de políticas y procedimientos para la preservación digital para abundar en el tema del desarrollo de una política.

Según la jerarquía de la política en su organización, la nueva documentación acerca del manejo de documentos de archivo en ambientes web y preservación de sitios web puede ser independiente, o bien, formar parte de una política más amplia. Además de los elementos descritos en el módulo 2 de política, la documentación sobre gestión y preservación de sitios web deberá incluir elementos o secciones adicionales. En general, la estructura de su política estará basada en la información contextual de su organización.

Gestión archivística

Todos los datos asociados con la preservación de sitios web deberán estar incluidos en las tablas de retención que regulan los documentos de archivo de la institución. Las páginas web deberán estar sujetas a la misma gestión archivística que se aplica a otros documentos de archivo electrónicos porque proporcionan evidencia de las actividades en línea de la organización. Además, si se cuenta con fechas de disposición en la gestión archivística, la organización se beneficiaría en cuanto a los costos asociados con el almacenamiento. Para asegurar la accesibilidad de datos a largo plazo es esencial que el medio de almacenamiento sea actualizado *(refreshing)* a intervalos regulares. Si la organización almacena cada iteración del sitio web indefinidamente entonces los costos asociados con los medios de almacenamiento se dispararán en el tiempo según el crecimiento de los datos acumulados.

Metadatos

Los metadatos son la clave para manejar de forma efectiva todos los documentos de archivo, incluyendo aquellos basados en actividades web. Los *Lineamientos para archivar recursos web de Australia*[5] describen los requisitos de metadatos requeridos para diferentes escenarios.

En documentos de archivo individuales en sitios web y en documentos de archivo basados en actividades web significa el uso de metadatos para describir:

- Fecha y hora de producción, y registro del documento de archivo en un sistema de gestión archivística.

[5] National Archives of Australia, *Archiving Web Resources.*

- Contexto organizacional.
- Formato de datos original.
- Uso que se hace del documento de archivo en el tiempo, incluyendo su colocación en un sitio web.
- Mandatos que regulan la producción, retención y disposición de documentos de archivo.
- Gestión de la historia del documento de archivo después de su producción, lo que incluye mantenimiento, preservación y disposición.

Para copias o instantáneas de todas las colecciones de recursos web, los metadatos deben incluir:

- Fecha y hora de captura
- Vínculos hacia el Identificador Uniforme de Recursos (URI por sus siglas en inglés), incluyendo información acerca de la versión y fecha del vínculo al URI especificado.[6]
- Detalles técnicos acerca del diseño del sitio web
- Detalles acerca de los programas o aplicaciones utilizados para producir recursos web
- Detalles acerca de las solicitudes (incluyendo los motores de búsqueda) que complementan los recursos web
- Detalles acerca de la solicitud requerida por el cliente para ver el recurso web.[7]

[6] Los lineamientos *Australian Guidelines for Archiving Web Resources* hacen distinciones entre un URI, un URL y un URN de este modo: el identificador universal de recursos (URI) es simplemente un mecanismo de vocabulario indexado (*namespace*) de propósito general; el localizador universal de recursos (URL) es una instancia del URI que es la dirección de algún recurso accesible mediante un protocolo como el HTTP; el nombre universal de recurso (URN) es una instancia de URI, que a diferencia de un URL que es frágil, está garantizado a permanecer disponible (Jon Udell, *Practical Internet Groupware*, Sebastapol, CA: O'Reilly, 1999, p. 471.)

[7] *Archiving Web Resources: Guidelines for Keeping Records of Web*, pp. 17-18.

Se recomienda que la auditoría de metadatos se lleve a cabo en el momento en que se emprenda un plan o programa de preservación de sitios web. Esto ase-gurará que los recursos capturados contienen los metadatos suficientes para preser-var de forma efectiva la exactitud, autenticidad, fiabilidad, accesibilidad y disposición de los recursos y que se permita el acceso así como que las actividades de preservación se lleven a cabo.

Véase, para más información, el módulo 4, Breviario de metadatos.

Gestión de derechos y propiedad intelectual

Las cuestiones relacionadas con los derechos de propiedad intelectual, tales como los derechos de copia (*copyrights*) y los derechos morales tienen un importante impacto en cualquier proceso de preservación digital.

> Los temas de derechos de propiedad intelectual en materiales digitales son más complejos y significativos que aquellos en medios tradicionales y si no son atendi-dos pueden obstruir e inclusive impedir las actividades de preservación.[7]

No se trata tan sólo del contenido, sino de los programas o apli-caciones asociados, los cuales pueden estar sujetos a derechos de propiedad intelectual.

[7] Jones Maggie y Neil Beagrie, *Preservation Management of Digital Materials*, p. 32.

> Simplemente el copiado (*refreshing*) de materiales digitales a otro medio, el encapsulamiento del contenido y los programas para emulación o migración del contenido a un equipo o programas nuevos involucra actividades que pueden infringir los derechos de propiedad intelectual, a menos de que existan exenciones regulatorias o se hayan obtenido permisos específicos de los propietarios de los derechos.[9]

Debido a la naturaleza de los materiales digitales, las estrategias para la preservación y el acceso continuos necesitarán la migración de materiales a nuevas formas o a una emulación del ambiente operativo original. Tales actividades requieren permisos de los propietarios de los derechos para emprenderlas.

Un campo que podría convertirse potencialmente problemático es el de las leyes de derechos de autor. Según la Canadian Heritage Information Network (CHIN): "el derecho de autor protege la expresión de ideas que están fijas en cualquier forma de medio".[10]

Esto incluye varios componentes del sitio web, tales como las imágenes que aparecen así como el código de programación.

El derecho de autor protege la mayoría de las creaciones tales como trabajos literarios, teatrales y artísticos, grabaciones y trabajos audiovisuales. Las fotografías son consideradas trabajos artísticos. Los programas de computadora en forma de código subyacente en una página web también se reconocen como trabajos literarios y, por tanto, protegidos por el derecho de autor. Con excepción de los trabajos producidos como

[9] *Idem.*
[10] Pantalony, Rina Elster, *Protecting your Interests*, p.13.

obligaciones de un empleado o donde el derecho de autor ha sido cedido por escrito a alguien más, el autor del trabajo es el propietario del derecho de autor.[11]

Se deben asentar quiénes son los propietarios de derechos de autor y asegurar los permisos antes de iniciar un programa de preservación en web.

Desarrollo y capacitación del personal

La capacitación cuidadosamente diseñada para el personal y el desarrollo profesional continuo pueden jugar un papel clave en el manejo exitoso de cualquier programa de preservación. Todos aquellos responsables de la preservación digital deben contar con cierto grado de conocimiento sobre el tema. El desarrollo y capacitación del personal pueden ir desde mantenerse actualizado con la literatura y los nuevos desarrollos hasta participar en talleres o cursos de capacitación impartidos por varias instituciones, tales como asociaciones archivísticas e instituciones educativas.[12]

Descripción del recurso, documentación y acceso

Una forma de descripción de clasificación es esencial para manejar cualquier colección archivística y hacerla accesible a usuarios; lo mismo vale para una colección digital. Los grandes estándares para catalogación, tales como MARC 21 e ISAD(G), han sido aplicados exitosamente para la descripción de sitios web que se archivan. La catalogación y clasificación de materiales archivados permiten a los usuarios acceder a éstos.

[11] *Idem.*

[12] The Society of American Archivists es una institución que organiza numerosos talleres y seminarios vía la web. Para una vista de su programa de actividades actuales véase: http://saa.archivists.org/Scripts/4Disapi.dll/4DCGI/events/ConferenceList.html?Action=GetEvents.

Los recursos deben ser suministrados con documentación apropiada y suficiente para satisfacer los requisitos para uso informado por parte de los miembros de la comunidad de investigación. La documentación debe relacionarse, tanto con el contenido como con el formato técnico del recurso. La documentación debe proporcionar también información acerca del contexto en el cual los recursos fueron producidos y mantenidos antes de su preservación así como las características y vínculos del recurso digital y de otras fuentes de información relacionadas.

Plan de recuperación en caso de desastre

El desarrollo de un plan de recuperación en caso de desastre basado en principios sólidos debe contar con la aprobación de la dirección y debe poder activarse por el personal capacitado. Esto reducirá ampliamente la severidad del impacto hacia la organización en caso de desastre. El plan necesitará orientarse a la restauración tanto del contenido del archivo como de la infraestructura técnica y operacional requerida para el soporte. Los elementos que deberá incluir el plan son:

- Asegurar que el personal se capacite para actuar en caso de desastres.
- Hacer copias de archivos de recursos de datos cada vez que se compila una colección de materiales.
- Guardar copias y archivarlas en múltiples medios
- Almacenar copias archivadas dentro y fuera del sitio
- Obtener documentación completa de la infraestructura de equipo y programas de cómputo así como de los procedimientos operativos y manuales.
- Contar con copias de todos los programas de cómputo requeridos para operar los sistemas.

También es importante probar el plan para identificar cualquier cuestión que pueda haberse omitido antes de que un desastre ocurra. Esto también ayuda al personal a familiarizarse con esos procedimientos con anterioridad. Como con la mayoría de las políticas, se recomienda que el plan de recuperación en caso de desastre sea revisado cuando se produzcan cambios en sistemas y circunstancias.

Ejercicios:

- ¿Actualmente en su organización se cuenta con un plan en caso de desastres?
- Analice la forma en la cual los recursos de su organización podrían ajustarse a este plan.

Verificaciones de validación

Una vez que el sitio web se ha integrado y transferido al archivo histórico de la institución, se deben llevar a cabo verificaciones para asegurar que todas las partes trabajan como se debe. Las verificaciones incluyen, entre otras: teclear manualmente todos los hipervínculos; hacer clic al azar sobre los hipervínculos o emplear el uso de una aplicación de prueba de estos vínculos para ayudar a automatizar el proceso de verificación en el sitio web que se trabaja,[13] verificando que los archivos pueden ser leídos, están completos y correctos así como la funcionalidad dentro de éstos. Estas verificaciones deben hacerse cuando un sitio web se ha archivado, a fin asegurar que el contenido y las estructuras de los recursos de datos están intactos.

[13] Véanse como ejemplos: Link Checker Pro: http://www.link-checker-pro.com/; Site Audit: http://www.blossom.com/site_audit.html; Cyber Spyder Link Test: http://www.cyberspyder.com/cslnkts1.html; Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

Formatos de archivo

En cualquier programa de preservación de sitios web (como con todo programa de preservación digital), antes de emprender cada estrategia de acopio se recomienda que sean definidos los formatos de archivo que se aceptarán. La adopción de un solo formato asegura que los costos de sustentabilidad se reduzcan cuando el formato de archivo seleccionado es incorporado desde el inicio del proceso de producción de los documentos de archivo.

> Se ha convertido en práctica común en los repositorios de documentos de archivo digital, incluyendo archivos históricos, el hecho de aceptar ciertos formatos de archivos digitales para la preservación en el largo plazo y rechazar otros.[14]

Las encuestas de instituciones respecto de las especificaciones de formatos de archivo muestran que existe una plétora de definiciones, formatos aceptables o inaceptables, e iniciativas de preservación para formatos de archivos.[15] El *Diccionario para la Preservación de Metadatos* Premis ofrece la definición más útil:

> una estructura específica preestablecida para la organización de un archivo digital o de una cadena de *bits*. Esta estructura preestablecida incluye la forma en que están codificados los datos y la forma en la cual los *bits* son interpretados para producir texto, imágenes y sonido.[16]

Esto es importante porque debe hacerse énfasis en el hecho de que es esencial definir los formatos de archivo aceptables para un repositorio específico.

[14] Peters McLellan, Evelyn, *General Study 11 Final Report*, p. 2.
[15] *Idem.*
[16] *Idem.*

> Algunos tipos de codificación son sinónimo de formatos específicos de archivo; por ejemplo, la codificación mp3 es utilizada para codificar un formato de archivo mp3.[17]

Esto parece suficientemente simple de entender pero en la práctica puede ser complicado. Si tomamos por ejemplo los archivos de texto plano "éstos pueden tener formatos con codificaciones diferentes, ya que pueden ser codificados como ASCII, EBCDIC o Unicode, entre un gran número posible de variantes."[18] Si sólo el texto plano tiene tres tipos diferentes de codificación, es obvio que los archivos de imagen y música no son tan simples.

> La codificación puede ser problemática en un formato de archivo de audio y video debido a que la codificación óptima para almacenamiento y transmisión con frecuencia conlleva la compresión (remoción de *bits* de los archivos digitales para reducir su tamaño), lo cual con frecuencia entorpece los esfuerzos de preservación.[19]

Dificultades adicionales surgen en cuanto a los formatos de archivos:

> La cuestión de codificación es más complicada por el hecho de que TIFF, WAV y AVI así como otros formatos de imagen y audiovisuales comunes no son "formatos" de archivo propiamente dichos, sino que son "formatos envolventes" (también llamados formatos contenedores), los cuales están diseñados para combinar cadenas de *bits* de naturaleza distinta en un sólo archivo de computadora.[20]

[17] *Idem.*
[18] *Idem.*
[19] *Idem.*
[20] *Idem.*

La codificación, las combinaciones de compresión y cadenas de *bits*: todas complican la forma en que los formatos de archivo son preservados a largo plazo. Éstas también son razones del por qué muchas instituciones piden formatos abiertos que estén bien certificados para asegurar que haya suficiente documentación disponible para brindar a la institución colectora la posibilidad de preservar los documentos de archivo a largo plazo.

Adrian Brown, archivista de los Archivos Nacionales de Reino Unido, ha identificado criterios que se tienen que considerar cuando se seleccionan los formatos de producción de datos; entre éstos están:

- Ubicuidad.
- Asesoría o soporte.
- Confidencialidad.
- Calidad de la documentación.
- Estabilidad.
- Facilidad de identificación.
- Derechos de propiedad intelectual.
- Soporte de metadatos.
- Complejidad.
- Interoperabilidad.
- Viabilidad.
- Reusabilidad.

Aunque la investigación no recomienda en especial tipos actuales de archivos, es importante tener en mente estos criterios cuando se seleccionan los formatos de archivo.[21]

Es importante que las organizaciones productoras y preservadoras desarrollen una política que establezca los tipos de

[21] Brown, Adrian, *Selecting File Formats.*

formatos de archivo que son aceptables. Al establecer restricciones en cuanto a los formatos de archivo que la institución aceptará recibir y manejará se asegura que los formatos de archivo que se recopilen cumplen con los criterios establecidos mencionados anteriormente y que se sujetan a los estándares actuales. Si se almacenan "buenos" formatos de archivo, las dificultades en la preservación se reducirán, así como los costos.

> Muchos formatos son licencias, esto es, son propiedad de un dueño quien, por razones comerciales, no está dispuesto a proporcionar acceso a la documentación que soportan y puede requerir que se le pague una tarifa por su uso.[22]

Ésta es la razón por la cual la mayoría de los expertos recomiendan formatos de archivo de estándares abiertos; también por qué se han desarrollado muchos registros de formatos de archivo. Los registros existen para proporcionar información fiable y detallada acerca de los formatos de archivo. Ejemplos de registros de formatos de archivo incluyen a: PRONOM[23] y el Registro Global de Formatos Digitales (Global Digital Format Registry).[24] En abril de 2009 la iniciativa del Registro Global de Formatos unió esfuerzos con la iniciativa PRONOM de los Archivos Nacionales de Reino Unido bajo un nuevo nombre, el Registro Unificado de Formatos Digitales (UDFR por sus siglas en inglés). El UDFR mantendrá los requisitos

[22] Ross Harvey, *Preserving Digital Materials,* p. 141.

[23] PRONOM es un registro de formatos de archivo establecido por los Archivos Nacionales de Reino Unido para proporcionar y manejar información acerca de los formatos y aplicaciones de cómputo utilizados. El sitio web de PRONOM puede encontrarse en: www.nationalarchives.gov.uk/pronom.

[24] El Registro Global de Formatos Digitales fue también desarrollado para apoyar a la preservación digital. Disponible en: http://www.gdfr.info/

y casos de uso compilados por el Registro Global de Formatos Digitales y se integrará con la base de datos de aplicaciones y formatos de PRONOM.[25]

La organización responsable de acumular documentos de archivo puede ayudar a promover su producción sólida mediante la publicación de aquellos formatos de archivo que tienen mayores posibilidades de ser sustentables en un periodo y alentando a la producción de documentos de archivo que utilice estos formatos. Otra alternativa para la institución es convertir todos los materiales digitales preservados al formato de archivo seleccionado, una vez que el material ya está en el archivo histórico.

Medios de almacenamiento[26]

Sea cual fuere el método de captura, el sitio web necesita ser preservado y almacenado en un medio o soporte electrónico digital relativamente estable. En la actualidad, ningún medio electrónico digital puede ser considerado como archivístico debido a que los periodos en los que tales medios se han probado son todavía cortos, además de la cuestión de la obsolescencia tecnológica que es el resultado de cambios rápidos en este entorno.

Los dispositivos para almacenamiento están en desarrollo y cambio continuo. El actual "estado del arte" de este ambiente

[25] El Registro Unificado de Formatos está disponible en: http://www.udfr.org/

[26] En este trabajo presentamos el medio o soporte básico de almacenamiento para guardar documentos electrónicos. Es posible crear un repositorio para materiales digitales. Si usted requiere más información revise la norma ISO:14721:2003, comúnmente conocida como el estándar ISO:14721, también conocido como el "Modelo de Sistemas de Información de Archivo Abierto" (OAIS, por sus siglas en inglés) así como OCLC y NARA. *Trustworthy Repositories Audit & Certification: Criteria and Checklist* Version 1.0, 2007. Disponible en: http://www.crl.edu/PDF/trac.pdf

puede ser obsoleto en un periodo de tan sólo cinco años y es simplemente imposible mantenerlo por 20 años. Los medios electrónicos no son permanentes como es la creencia común. Los fabricantes aseguran satisfactoriamente tiempos de vida largos para sus medios;[27] pero la experiencia práctica sugiere que una visión realista para la vida de una cinta magnética puede ser de 15 años y para un disco compacto, 20 años; todo esto según la calidad del original, almacenamiento, manejo y uso. Aun si el tiempo de vida es mayor, el equipo para leerlo puede no estar disponible. Para muchos medios, una imperfección que aparece después de un tiempo puede hacer que todo el medio sea inutilizable.[28] Por tanto, cualquiera que sea el medio seleccionado para almacenamiento necesitará verificarse periódicamente y refrescarse para contrarrestar la pérdida de datos.[29]

Muchos factores afectan la longevidad de los medios electrónicos, incluyendo las condiciones de almacenamiento, la calidad y la composición de los productos utilizados debido a la disponibilidad de mejores materiales que aparecen con el tiempo. Por tanto, es muy difícil predecir la longevidad. El Instituto Canadiense de Conservación ha elaborado una tabla que proporciona estimaciones de longevidad para varios medios de almacenamiento (Tabla 1).

[27] En 1995 una investigación de Kodak en sus CD regrabables se reportó un periodo útil de 217 años bajo condiciones específicas. Disponible en: http://www.cd-info.com/archiving/kodak/index.html

[28] Linden Jim y Sean Martin, Richard Masters y Roderic Parker, *The large-scale archival storage of digital Objects*.

[29] Véase *The National Archives of the UK's Digital Preservation Guidance Note:* 2, "Selecting Storage Media for Digital Preservation", by Adrian Brown, Head of Digital Preservation Research, Agosto 2008. Disponible en: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Tabla 1 Longevidad pronosticada de medios electrónicos[30]

Tipo de medio	Longevidad pronosticada
Discos magnéticos	
Discos duros	2–5 años
Disquetes flexibles	5–15 años
Cintas magnéticas	
Digitales	5–10 años
Analógicas	10–30 años
Discos ópticos	
CD-RW, DVD-RW, DVD+RW	5–10 años
CD-R (cianina y colorantes azoicos)	5–10 años
Audio CD, DVD, cine	10–50 años
CD-R (colorante de ftalocianina, capa metálica de plata)	5–10 años
DVD-R, DVD+R	10–50 años
CD-R (colorante de ftalocianina, capa metálica de oro)	>100 años
Otros discos ópticos	
MO, WORM, etc.	¿10–25 años?
Dispositivos tipo flash	desconocido

Por tanto, se recomienda que el sitio web archivado sea almacenado en varios soportes, por ejemplo en un disco duro y en un DVD-R, y guardado en los archivos para contrarrestar estos problemas de almacenamiento y ayudar a asegurar el acceso de los datos almacenados a largo plazo.

[30] Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives.*

En la determinación del tipo de medio de almacenamiento para almacenar materiales digitales es necesario considerar ciertos factores, que incluyen, longevidad, capacidad, viabilidad, costos y sustentabilidad, según lo documenta Adrian Brown, de los Archivos Nacionales de Reino Unido.[31] Brown expone en una tabla de puntuación la comparación entre los tipos de soportes o medios comunes (Tabla 2).

Tabla 2 Comparación entre los tipos de soportes o medios comunes

Medio	CD-R	DVD-R	Disco Duro	Memorias y tarjetas flash (Dispositivos USB)	Cinta de Almacenamiento de datos (*Linear Tape Open* LTO)
Longevidad	3	3	2	1	3
Capacidad	1	3	3	2	3
Viabilidad	2	2	2	1	3
Obsolescencia	1	2	2	2	2
Costo	3	3	1	3	3
Susceptibilidad	1	1	3	1	3
Total	11	14	13	10	17

Según esta tabla, las soluciones con puntajes más altos son la cinta de almacenamiento LTO y el DVD-R, junto con la opción de disco duro como ua tercera opción. Brown aconseja:

> En situaciones donde las copias múltiples de datos son almacenadas en medios separados, puede ser ventajoso usar diferentes tipos de medios para cada copia, utilizando preferentemente diferentes tecnologías base (por ejemplo magnética y óptica). Esto reduce en general la

[31] The National Archives, *Digital Preservation Guidance Note 2: Selecting Storage Media for Long-Term Preservation*, agosto 2008. Disponible en: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

> dependencia a la tecnología de los datos almacenados. Donde se usa el mismo tipo de medio para copias múltiples, se deberán usar diferentes marcas o lotes en cada caso a fin de minimizar los riesgos de pérdida de datos debido a problemas con fabricantes o lotes específicos.

Joe Iraci, del Instituto Canadiense de Conservación, aporta comentarios adicionales respecto de las diferencias del medio de almacenamiento. En cuanto al uso de medios ópticos para almacenamiento, Iraci señala:

> el tipo de disco utilizado y la forma como es grabado impacta enormemente en la longevidad. [Él destaca] que las cintas digitales tienen corta duración y necesitan ser migradas/refrescadas cada 5-10 años [y advierte que] los discos duros no son para el almacenamiento en el largo plazo ya que los datos necesitan moverse a un nuevo disco duro cada dos a cinco años.

y nos recuerda "seguir con tecnologías ya generalizadas y evitar nuevas tales como discos Blue-Ray, almacenamiento holográfico [y] memorias flash". Iraci también señala que "con todos los medios digitales, los respaldos son críticos a fin de evitar una pérdida repentina de información"[32]

Las investigaciones, como las realizadas por Adrian Brown y el Instituto de Conservación Canadiense, son invaluables cuando hay que decidir el medio para el almacenamiento institucional de documentos de archivo. Es claro que se debería elegir una variedad de medios y que aún con el almacenamiento y manejo correctos el medio debe ser verificado y actualizado regularmente.

[32] Correo electrónico de Joe Iraci a Randy Preston, mayo 20, 2009.

Normas

Cierto número de estándares están relacionados con el archivado de sitios web. HTML y XML son el núcleo de tecnologías reconocidas como estándares en la forma de las recomendaciones del W3C.[33] Existen dos estándares en el área de gestión de documentos de archivos, ISO 15489-1/2:2001, el cual establece los estándares para la práctica de la gestión archivística e ISO 23081-1:2006, que establece estándares para la gestión de metadatos para documentos de archivo.

ISO 14721:2003 es la norma que define los requisitos fundamentales para un sistema de preservación digital. Más conocido como el Modelo de Referencia para un Sistema Abierto de Información Archivística (OAIS por sus siglas en inglés), sus conceptos y metodología han sido ampliamente adoptados internacionalmente; también es la base para un esquema de certificación de repositorios digitales de confianza.

Véase Módulo 1 Un esquema para la preservación digital, para abundar en el modelo de referencia OAIS.

En la ISO 19005-1: 2005 o el estándar PDF/A se ha abordado la necesidad de formatos de archivo digitales abiertos. Este estándar es

> un formato de archivo basado en PDF, conocido como PDF/A, el cual ofrece un mecanismo para representar

[32] W3C, o World Wide Web Consortium o el Consorcio de la web mundial es un grupo internacional donde las organizaciones miembros, personal de tiempo completo y el público trabajan juntos para desarrollar estándares para la web.

documentos electrónicos de tal forma que preserva su apariencia visual al paso del tiempo, con independencia a las herramientas y sistemas utilizadas para producir, almacenar o presentar los archivos.[34]

Mantener documentos de archivo basados en la web a través del tiempo

Asegurar la accesibilidad de los materiales basados en la web al paso del tiempo plantea las mismas cuestiones de accesibilidad que en el caso de otros documentos de archivo electrónicos. Existen pasos que pueden ser tomados para mitigar estas cuestiones, incluso el asegurar que los materiales son manejados cuidadosamente (desde mantener la confiabilidad de los documentos de archivo en la web hasta identificar y mitigar la gestión de riesgos), tales como planes para la obsolescencia, uso de estándares ampliamente sustentados, implementación de medidas de seguridad para proteger contra alteraciones deliberadas o accidentales y asegurar el control ambiental y supervisión. La mayoría de estos pasos ha sido analizado en la primera parte de este documento, pero es prudente insistir sobre su importancia.

Administración cuidadosa que puede incluir: mantener matrices (*masters*) y almacenarlas en una ubicación separada; implementar el uso de XHTML y evitar el uso de etiquetas no estandarizadas de HTML; refrescar el medio de almacenamiento periódicamente; llevar a cabo verificaciones al azar para asegurar la accesibilidad a los datos.

Planeación en caso de obsolescencia: planear la obsolescencia al asegurar que los documentos de archivo pueden ser copiados,

[34] ISO 19005-1 – Administración de Documentos – Formato de archivo para la preservación en el largo plazo de un documento electrónico – Parte 1: Uso de PDF 1.4 (PDF/A-1).

refrescados o migrados. Cualquier actividad de preservación como las mencionadas deberá estar documentada en los metadatos de gestión archivística, incluyendo toda pérdida de funcionalidad, contenido o apariencia.

Uso de normas y estándares: es importante el uso de normas y estándares.

Medidas de seguridad como medio para proteger datos. Es importante tomar medidas de seguridad en la preservación de la web que protejan los datos de alteraciones deliberadas o accidentales. Las medidas pueden ser tan simples como mantener los datos preservados en un ambiente seguro, con acceso controlado sólo a personas autorizadas que podrán leerlos pero no modificarlos.

Control ambiental y monitoreo. La mejor práctica establece que los medios almacenados deben ser mantenidos a niveles óptimos de temperatura y humedad; protegidos y alejados de campos magnéticos; se debe contar con unidades de filtración de aire para protegerlos contra impurezas; prohibir el consumo de alimentos en el área de almacenamiento; y establecer medidas en caso de desastre.

Véase la *Guía* NARA *sobre Manejo de Documentos de Archivo en la Web*, disponible en: http://www.archives.gov/records-mgmt/policy/managing-Web-records-index.html

Métodos de captura de sitios web/herramientas

Existen dos tipos de sitios web: estáticos y dinámicos. Un sitio web estático está compuesto de una serie de páginas construidas previamente, que están vinculadas desde al menos otra página. Un sitio web dinámico genera páginas sobre la marcha a partir de pequeños elementos de un contenido que puede estar guardado en una base de datos, recopilados desde fuentes externas e insertados en la página web, o generado mediante programas que responden de forma diferente según factores como la fecha u hora en la que se accedió a la página web.

Actualmente existen tres opciones disponibles para capturar sitios web. Los métodos para captura dependen de qué tanta información desea preservar la institución que lo conservará. Esta información incluye funcionalidad, metadatos y el grado de autenticidad, fiabilidad y exactitud que la institución desea preservar. Las tres opciones son: transferencia directa, recolección remota y sitio web espejo.

Transferencia directa

La única forma para recrear completamente un sitio web en un ambiente de preservación es a través de la transferencia directa de datos. Ésta opera mediante la adquisición de una copia de los datos directamente desde la fuente original. Esto requiere acceso directo al servidor primigenio de alojamiento de la web. Así, la transferencia directa consiste en copiar los archivos seleccionados desde el servidor y transferirlos a la institución que los recopila. Para garantizar la funcionalidad continua es necesario llevar a cabo ajustes menores al sitio preservado.[34] Para asegurar

[34] Por ejemplo, los hipervínculos dentro de un sitio archivado pueden necesitar ser ajustados desde vínculos absolutos a vínculos relativos y el motor de búsqueda apropiado (el utilizado en el ambiente original) debe ser instalado en un nuevo ambiente para asegurar que también la funcionalidad de búsqueda es preservada. Para una explicación más amplia véase: Adrian Brown, *Archiving Websites*.

que el sitio web preservado es lo más auténtico posible, será necesario implementar una recreación del ambiente técnico en el cual reside el sitio web en el entorno archivístico. Esto significa que la base de datos o el sistema de administración de contenidos requerirán ser instalados en el ambiente archivístico, junto con el servidor de web y los programas del motor de búsqueda necesarios. La transferencia directa es el único método que toma en consideración la naturaleza dinámica de un sitio web y es la única forma para preservar formas posibles de datos generados dinámicamente. Sin embargo, la implementación y soporte de tal método requerirá de la disponibilidad de personal con habilidades técnicas para instalar y mantener el sistema.

Recolección remota

La solución de recolección remota ofrece tres opciones: rastreo directo automatizado del sitio web, rastreo de imágenes "instantáneas" con bitácoras adicionales mantenidas por el archivista para respaldar los datos extraídos en la imagen, y la tercerización del proceso. Los métodos por recolección remota como opción deben emplearse con la advertencia de que no capturan la totalidad de las posibilidades de la página en la web, aunque pudieran generarse por solicitud del usuario cuando el sitio identificado para su captura es un sitio dinámico con una base de datos subyacente (*back end*) utilizada para albergar la información generada sobre la marcha. También, el uso de este método puede tener como resultado la presencia de hipervínculos rotos dentro del ambiente de datos copiados, debido a que las páginas pueden contener vínculos a contenido que necesita ser generado sobre la marcha para presentarse al usuario. Otra pérdida de datos que podría ocurrir puede ser la pérdida de gráficas y el diseño de plantillas.

Una instantánea de un sitio web usualmente implica la creación de una copia completa y exacta del sitio de una organización en

un momento específico. Una instantánea deberá incluir todos los aspectos del sitio web para asegurar que un sitio funcional completo puede ser recreado. Una instantánea deberá incluir guiones, programas, complementos (*plugins*), y las aplicaciones de búsqueda, componentes que en conjunto hacen que la instantánea sea completamente funcional.

Un estándar para rastreo de web debe hacerse utilizando un programa de recolección de sitio web de código abierto, como las herramientas gratuitas GNU Wget o Heritrix. Las ventajas de un rastreador de código abierto para archivo de sitio web son las de no ser propietario y, por tanto, no se incurre en erogaciones financieras. Un rastreador automatizado de web puede coleccionar datos tan frecuentemente como lo desee la institución; inicialmente el rastreador podría establecerse para rastrear todo el sitio, posteriormente los rastreos subsecuentes pueden coleccionar datos sólo de páginas que hayan sido actualizadas a partir del rastreo previo.

Para preservar instantáneas de un sitio web en cierto momento, la institución necesita rastrear el sitio sólo una o dos veces al año. Sin embargo, con esta frecuencia, obviamente no se captura cada cambio hecho y se pueden perder algunas de las actividades documentadas que se presenten. El rastreador de web puede ser implementado para realizar rastreos poco frecuentes del sitio. Así, se toman "copias instantáneas" del sitio web como un todo (asegurando que la funcionalidad de los vínculos internos no sea destruida y que se mantienen) y mientras tanto, para asegurar que la evidencia necesaria se captura, se lleva una bitácora de cambios que determina cuándo y cómo los documentos o las páginas web son removidos, reemplazados o actualizados. Si los propósitos de rendición de cuentas y mantenimiento son importantes a fin de que los documentos de archivo del contenido del sitio web y sus cambios sean realizados y mantenidos, entonces ésta

es una opción viable y barata.[36] Una vez más, los metadatos son la clave para manejar todos los documentos de archivo de forma efectiva, incluyendo aquéllos basados en la actividad web (véase el Módulo sobre metadatos).

Existe la opción de tercerizar o subcontratar con base en una tarifa, la captura y almacenamiento de sitios web. Los servicios tales como Web Archiving System (WAS) desarrollado por la Biblioteca Digital de California y el Proyecto Archive-It operado por el Internet Archive proporcionan servicios de captura y almacenamiento a organizaciones que desean preservar sus sitios web. Es importante mencionar que con esta opción, los datos almacenados hospedados en estos servicios pueden estar distribuidos por todo el planeta y sujetos a una variedad de leyes y reglamentos jurisdiccionales. Por ello, antes de emprender un proyecto de este tipo es primordial estar perfectamente al tanto de su esquema legislativo y regulatorio respecto de protección de datos y acceso a la información.

Sitio web espejo

El sitio web espejo es una opción de copiado, pero no captura los metadatos asociados necesarios para preservar de forma efectiva el contenido digital del sitio web. Un espejo es una copia exacta de un conjunto de datos. Esencialmente opera como una "copia impresa" pero en forma digital. El espejo de sitios se lleva a cabo por varias razones, una de éstas para preservar un sitio o una página web.

Como se señaló anteriormente, no capturar los metadatos asociados con cada archivo de la página web es una buena opción

[36] El rastreo de web con opción de registro fue investigado utilizando la obra *Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government* de los Archivos Nacionales de Australia. Es un documento de gestión documental de gobierno publicado en marzo de 2001 y puede verse en: http://www.naa.gov.au/Images/archWeb_guide_tcm2-903.pdf

cuando el archivo histórico sólo desea preservar evidencia del sitio web en cuestión. Esta solución debe ser entendida con la previsión de que no hay metadatos de evidencia de los documentos de archivo que puedan aparecer en el sitio. Por lo tanto, no se recomienda si el archivo histórico que los colecciona desea preservar evidencia de los documentos de archivo que aparecen en el sitio web.

Herramientas de captura de web

El rastreador de código abierto HTTrack ha sido utilizado de forma efectiva en otras instituciones archivísticas;[37] es gratuito y es una herramienta para navegar sin conexión y de fácil uso. Permite al usuario bajar un sitio web desde Internet a un directorio local, construir de forma recurrente todos los directorios, copiar HTML, imágenes y otros archivos desde el servidor al directorio local. HTTrack arregla la estructura de enlaces relativos del sitio original. Permite a los usuarios simplemente abrir una página con el sitio web "reflejado" en su rastreador y buscar el sitio de vínculo a vínculo, como si se viera en línea.[38] Los archivistas que buscan preservar contenidos de web en ambientes de Microsoft Windows han usado este recolector con éxito.

Véase "Practical E-Records" para una revisión de las herramientas HTTrack, GNU Wget, Heritrix y Web Archiving Service. Disponible en: http://e-records.chrisprom.com/?tag=Website-harvesting. Todo esto fue revisado con base en los siguientes criterios: instalación/configuración/plataformas o portadas/funcionalidad/fiabilidad; usabilidad; escalabilidad, documentación; interoperabilidad/metadatos de soporte; flexibilidad/personalización; licencia/soporte/sustentabilidad/comunidad.

[37] Correo electrónico enviado a la lista *Management & Preservation of Electronic Records Listserv*, abril 3, 2009, por parte del Archivista de Documentos de Archivo Electrónicos del Departamento de Archivos y Bibliotecas de Kentucky.

[38] Véase el sitio web de HTTrack para mayor información en: http://www.httrack.com/

Además, la herramienta de captura de web de Adobe (Web capture) convierte las páginas web en archivos de PDF para crear versiones PDF de la página web. Es simple de usar y, por tanto, fácil de enseñar al personal. Es posible capturar todo un sitio web usando esta herramienta. No sólo todos los vínculos continúan operando en el PDF, también se puede vincular el contenido local dentro del PDF cuando es pertinente, de tal forma que se puede navegar en el sitio sin conexión. Web Capture puede ser solicitada directamente desde la barra de herramientas de Acrobat en Internet Explorer en Windows o a través de la aplicación de Adobe Acrobat 9 en plataformas Windows y Mac.

La herramienta es de uso fácil, captura varios niveles de vínculos dentro del sitio, tiene un sello que estampa fecha y hora en las páginas capturadas, y Adobe asegura compatibilidad con versiones anteriores. Sin embargo, no hay metadatos capturados; reproduce un documento plano en PDF, lo que significa que no es posible remover una porción de una página para imprimir, por ejemplo sólo una fotografía, por lo que se tiene que imprimir la páginacompleta; el sitio web completo es capturado cada vez, y la herramienta convierte el sitio web a PDF simple en lugar de un PDF/A.

Adobe liberó la herramienta de captura de sitios web en 2008 y es una herramienta extremadamente simple para implementar y usar. Adobe tiene una buena reputación y una historia de soporte para el cliente, y trata de asegurar que cada producto nuevo sea compatible retrospectivamente con las versiones previas.

Una herramienta más para archivar sitios web (construidos dinámicamente con una base de datos subyacente para hospedar la información generada por el usuario sobre la marcha), es la de archivar la base de datos. La técnica es apenas incipiente pero vale la pena describirla ya que es una herramienta que puede ser

utilizada para paliar problemas asociados con el archivado de sitios web dinámicos mediante el uso de métodos de sitio web estáticos.

Brown describe el proceso de archivar sitios web por medio de bases de datos en tres etapas. Primero el repositorio define el modelo estándar de datos y formato para las bases de datos a archivar. Luego, la base de datos fuente es convertida al formato estándar, y finalmente, se proporciona una interfaz de acceso estándar para las bases de datos archivadas.[39]

Los Archivos Federales de Suiza han desarrollado un formato basado en XML que permite la preservación en el largo plazo de contenidos de bases de datos relacionales de manera independiente de los programas o aplicaciones. El formato tiene una larga historia desde principio de los años noventa. En mayo de 2008 fue aceptado como el formato del proyecto europeo PLANETS para archivar bases de datos relacionales. El formato se conoce como Software Independent Archiving of Relational Databases (SIARD). Preserva datos de contenido y metadatos así como las relaciones en un formato que cumple con normas ISO. Un artículo informativo publicado en octubre de 2008 por la organización Digital Preservation Europe, Database Preservation: The International Challenge and the Swiss Solution describe el proceso de SIARD.[40]

[39] Brown, *Archiving Websites*, p. 59.

[40] Según el artículo informativo publicado en octubre 2008 por Digital Preservation Europe, "Database Preservation: The International Challenge and the Swiss Solution" (http://www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf), "El uso de estándares ISO extensamente aceptados asegura en gran medida que los datos almacenados puedan ser accedidos en el futuro. Basados en esta presunción los documentos de archivo SIARD, tanto datos como metadatos se recaban automáticamente en formatos de la norma ISO: SQL 1999 Unicode, y el más importante de éstos: XML1.0. Para asegurar la normalización SIARD convierte todas las bases de datos propietarias en el equivalente al conjunto de caracteres Unicode. Además SIARD no archiva sinónimos ya que no son parte del SQL:1999 estandarizado. Apegarse a los estándares es una regla de hierro".

Un archivo SIARD es un contenedor ZIP estructurado no comprimido (ZIP-64 estándar), que permite prácticamente cualquier tamaño de archivo. Contiene dos carpetas: "encabezado" y "contenido". La carpeta encabezado almacena el contexto de la base de datos y metadatos. Un solo archivo, metadata.xml, asegura que podamos entender los antecedentes técnicos así como los de contexto de la base de datos. En términos técnicos, SIARD registra el nivel más alto (la base de datos) el identificador, la versión de formato, el código de resumen de mensaje de la terminal PC que archiva (verifica los datos primarios de integridad), etcétera. Sobre el nivel de esquema, SIARD almacena listas de tablas, vistas y rutinas. A nivel de tabla, SIARD registra las limitantes y los "disparadores" de eventos. Más a fondo, en el nivel de columna, SIARD también específica el tipo de lenguaje SQL en uso, los nombres de objetos de gran volumen, y lo más importante de todo: claves externas y claves candidatas con datos interreferenciales; es decir, las relaciones. Al mismo tiempo SIARD contextualiza los datos. En el nivel de base de datos nos permite registrar o aumentar (con la suite SIARD) información sobre la procedencia del archivo, descripción, usuario, etcétra. En niveles más bajos nos permite mantener detalles de las tablas y nombres de las columnas y el contenido. Toda esta información descriptiva hace que la base de datos sea comprensible para los futuros usuarios, tanto en términos contextuales como técnicos.

La segunda carpeta, la de contenido, almacena los datos primarios. Los datos son archivados de acuerdo con la estructura de la base de datos. Para cada esquema, SIARD genera automáticamente una carpeta (cuadro 1, cuadro 2, etcétera.) que contiene la serie de tablas correspondientes como subcarpetas (tabla 1, tabla 2, etcétera). Los datos mismos son almacenados en archivos XML (por ejemplo, Tabla1.xml). Esta definición de esquema refleja el esquema de metadatos de la tabla SQL, y especifica que la tabla es almacenada como una cadena de líneas que

abarcan una secuencia de entradas de columna con tipos XML diferentes. Los grandes conjuntos u objetos binarios o de caracteres (BLOB y CLOB que contienen todos los tipos de información) también son archivados; son almacenados en carpetas generadas automáticamente (por ejemplo, lob1, lob2, etcétera) ya sea en archivos TXT o BIN (record1.text, o record1.bin, etcétera.[41]

SIARD es también un formato abierto, lo cual significaría que la organización que colecciona podría, de hecho, archivar la base de datos sin costos adicionales posibles por obtener una licencia al propietario del sistema administrador de contenidos requerido si el método de transferencia directa de captura es empleado.

De la descripción anterior de SIARD es claro que la institución que colecciona necesitará contar desde el inicio con una persona con preparación tecnológica para implementar exitosamente la suite o ensamble de programas SIARD.

En estos momentos es incierto si el ensamble SIARD está disponible para uso público. En la presentación de Jean-Marc Comment, representante de los Archivos Federales de Suiza, durante el 16 Congreso Internacional de Archivos, en julio de 2008[42] se mencionó que las herramientas de SIARD estarían disponibles en un futuro próximo por parte de los Archivos Federales de Suiza. Hasta el 12 de octubre 2009, en el sitio web[43] no aparecía nada respecto de estas herramientas. Debido a la duda sobre la disponibilidad, la suite SIARD no ha sido incluida como una opción de preservación en este trabajo. Sin embargo, es una opción interesante que puede seguirse una vez que su disponibilidad sea real.

[41] Para una análisis más completo del formato SIARD, véase *SIARD Format Description*, disponible para descarga en: http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=de

[42] Para ver toda la presentación, véase: http://www.planets-project.eu/docs/presentations/ICA2008_Comment_SIARD.pdf

[43] Sitio web de los Archivos Federales de Suiza: http://www.bar.admin.ch/index.html?lang=en

Plan de acción general para la preservación de un sitio web

Aunque no existe una solución genérica para los planes de preservación de sitios web, existen ciertos elementos que son universales para todos los planes y programas. Cuando se ha identificado la estrategia más apropiada debe ser seleccionado un equipo que incluya gestores de documentos o archivistas, administradores de sitios web, administradores de comunicación y personal de tecnologías de la información. El equipo puede desarrollar un plan de acción general que incluya políticas y procedimientos adecuados a sus necesidades.

A continuación se presenta un plan de acción general para la preservación de un sitio web que puede ser adaptado a las diferentes necesidades de una institución:

1. Identificar los requisitos de gestión archivística para la actividad basada en web.
2. Determinar si el sistema existente satisface los requisitos señalados o si es necesario diseñar e implementar uno nuevo o mejorar el actual.
3. Elevar el perfil y la conciencia general dentro de la organización de las responsabilidades de gestión archivística de todo el personal.
4. Llevar a cabo una evaluación de riesgos para determinar el nivel de riesgo aceptable planteado.
5. Desarrollar una política global para la preservación del sitio web (o una política de preservación de documentos de archivo que incluya la preservación del sitio web).
    a. Desarrollar una política de recopilación (incluye una política de selección).[44]

[44] Tanto la política como las listas de recolección deberán ser revisadas periódicamente para aumentar o quitar recursos.

b. Desarrollar una política de selección.
    i. Definición del contexto.
    ii. Métodos de selección.
    iii. Criterios de selección.
        (1) Valoración de contenido.
        (2) Extensión.[45]
    iv. Lista de recolección.
        (1) Selección de los recursos del sitio web a recuperar.
    v. Definiciones de límites.
        (1) Determinar URL o nombre de dominio.[46]
        (2) Parámetros.[47]
    vi. Definir el método de recolección.
    vii. Determinar tiempos y frecuencia de recolección –incluyendo la metodología de evaluación de riesgos.[48]

[45] Es necesario establecer criterios para determinar la extensión de los recursos seleccionados; esto es, si se coleccionarán o no vínculos externos.

[46] Si un solo recurso de web, tal como una página o documento va a ser coleccionado de forma aislada, entonces la lista de recolección simplemente necesita especificar el URL de ese recurso. Si todo el sitio web ha sido seleccionado, usualmente será definido por el nombre de dominio.

[47] Los parámetros definen el número de niveles de la estructura de directorio que será coleccionada y si los vínculos externos serán o no seguidos y si así fuera a qué profundidad.

[48] La Universidad de Cornell ha desarrollado una metodología para evaluar y mitigar riesgos para recursos de web vivos: El Proyecto Virtual Remote Control (VRC), de Cornell, está disponible en: Resumen: La metodología VRC para el análisis de riesgos sigue un proceso de seis pasos que inicia con la identificación y la evaluación de sitios web, facilita la evaluación del nivel de riesgos del sitio y la estrategia de construcción, e inicia una respuesta subsecuente. El catálogo VRC busca automatizar este proceso lo más posible, pero permite el control humano. La estabilidad del sitio web es medida a varios niveles de riesgo que pueden ser deducidos mediante la supervisión conforme avanza el tiempo respecto de su implementación (por ejemplo, arreglo u orden de HTML) y su estructura de hipervínculo, así como desde los metadatos del servidor de web (p. ej., programas del servidor, tiempo de respuesta). Si el sitio web se encuentra en un riesgo alto puede ser necesario establecer contacto con el dueño del sitio. VRC planea entonces establecer recomendaciones en los 'lineamientos para la preservabilidad de contenidos de web'. Como último recurso para recursos de web en riesgo éstos pueden ser también cosechados y preservados para evitar su pérdida. Resumen de ERPANET.

(1) Influencia del ciclo de vida del recurso de web.
(2) Tasa o rango de cambio de contenido.
(3) Actualidad y significado.

viii. Definir almacenamiento para los activos digitales.[49]

6. Implementar la política.
7. Documentar procedimientos y procesos para asegurar que las estrategias se llevan a cabo.
8. Iniciar el programa de preservación del sitio web.
9. Realizar verificaciones en los datos capturados y almacenados.[50]
10. Revisar la política y los objetivos de valoración de forma frecuente.

[49] Para asegurar la accesibilidad en el largo plazo, es esencial que los medios de almacenamiento sean refrescados periódicamente; la acción de refrescar el medio de almacenamiento deberá ser construida dentro de la política general de documentos de archivo electrónicos como se vio en los pasos de los planes de acción.

[50] Una vez que el sitio web ha sido capturado y transferido al ambiente del archivo histórico, se deben llevar a cabo verificaciones para asegurar que todas las partes del sitio están trabajando como deberían. Las verificaciones incluyen, entre otras: teclear manualmente todos los hipervínculos; teclear aleatoriamente algunos hipervínculos, o emplear el uso de una aplicación de prueba de existencia de vínculos como ayuda a automatizar el proceso de verificación. Ejemplos de aplicaciones de prueba de vínculos son: Link Checker Pro: http://www.link-checker-pro.com/ Site Audit: http://www.blossom.com/site_audit.html Cyber Spyder Link Test: http://www.cyberspyder.com/cslnkts1.html y Link Sleuth: http://home.snafu.de/tilman/xenulink.html.

Estudio de caso: desarrollo de un plan de preservación de un sitio web para una asociación académica en una institución académica

En esta sección se analiza un estudio de caso del desarrollo de un plan de preservación de un sitio web hipotético para una asociación de estudiantes dentro de una gran institución académica. Este estudio ofrece un ejemplo acerca de cómo identificar requisitos y desarrollar un plan de preservación de un sitio web organizacional. La meta del estudio es desarrollar estrategias para ejercer un mayor control de las modificaciones del sitio web de la asociación y para la preservación a largo plazo de varias iteraciones.

Antecedentes sobre la organización

La Asociación de Estudiantes de la Universidad Acme (ASUA)[51] está ubicada en el campus de la casa de estudios. La asociación cuenta con 30,000 miembros, estudiantes del campus principal y de otros externos a la Universidad. La ASUA opera como una asociación independiente sin fines de lucro. Su propósito es supervisar los servicios a estudiantes (por ejemplo, tutorías, búsqueda de empleos, etcétera), negocios y clubes. La ASUA es el centro de gestión archivística y archivo histórico para la sociedad "Alma Mater".

El archivista de ASUA buscó estrategias para la preservación en el largo plazo de un sitio web que cambia frecuentemente; estaba interesado en desarrollar estrategias para ejercer un mayor control sobre las modificaciones al sitio y para la preservación a largo plazo de iteraciones. El resultado del estudio es un plan de acción que concibe estrategias para el control sobre el sitio web y su preservación a largo plazo.

[51] Una universidad y una sociedad ficticias.

Los retos

La ASUA tiene recursos limitados (experiencia tecnológica, personal, tiempo, recursos financieros, etcétera) para desarrollar y sustentar una estrategia para la preservación en el largo plazo de su sitio web.

El proceso de desarrollo

El sitio web de ASUA fue identificado como el cuerpo de material digital para el cual sería desarrollado un plan de preservación. Los datos coleccionados fueron acerca del contexto de la institución y sus limitantes, el cuerpo específico de material, sus formas documentales, las restricciones tecnológicas y el sentido funcional y cultural de los materiales.

Identificar el contexto

Se compiló información relacionada con la institución, sus documentos de archivo y sus operaciones a través de una aproximación etnográfica. Se llevaron a cabo varias entrevistas y observaciones con el archivista de la Asociación, el gerente de comunicaciones, el editor del sitio web y el administrador de tecnologías de la información, para obtener como resultado el análisis contextual y diplomático así como información de los documentos de archivo producidos por la ASUA y una perspectiva cultural de los responsables del sitio web.

Método para desarrollar un procedimiento de mantenimiento

Se emprendió el desarrollo de un documento de procedimientos que delinea cómo el sitio web de AUSA va a mantenerse. Este documento establece los procedimientos para el mantenimiento del sitio web

y los procedimientos que contienen los criterios que se seguirán para lo que puede residir en el sitio. Tales procedimientos gobernarán el contenido del sitio tomando en consideración las restricciones que pueden ser establecidas en el contenido, en parte debido a la necesidad de adherirse a requisitos y normas administrativas, jurídicas y legales. Contar con tales procedimientos también asegura una valoración precisa y exhaustiva en el futuro.

Valorar el sitio web

Se llevó a cabo la valoración del sitio web de ASUA para establecer lo que sería preservado. Se hicieron cuatro preguntas clave: 1) ¿Qué capturar? 2) ¿Con qué frecuencia capturar? 3) ¿Qué tanto capturar? y 4) ¿Por cuánto tiempo preservar lo que se captura?

Investigación sobre la mejor estrategia para preservar el contenido del sitio web

Investigar e identificar las opciones tecnológicas que cumplan los objetivos de valoración de ASUA y sus restricciones tecnológicas, financieras y humanas; determinar los costos de los recursos para implementar las opciones tecnológicas identificadas. La investigación buscó reconocer métodos para la preservación del sitio que han sido instrumentados exitosamente en otras organizaciones parecidas, así como averiguar el conocimiento construido por grandes organizaciones. Además se investigó sobre los métodos que no han sido instrumentados actualmente.

Muchas grandes organizaciones han contribuido en el desarrollo de métodos para la captura y preservación del sitio web, y las mismas fueron investigadas en cuanto a metodologías probadas. Dentro de las grandes organizaciones que actual-

mente preservan sitios web útiles para este proyecto se tienen a la Biblioteca del Congreso, el Internet Archive, los Archivos Nacionales de Reino Unido y los Archivos Nacionales de Australia. Cada una de estas instituciones fue de gran ayuda para entender los componentes que se requieren integrar en una estrategia de preservación. La mayoría de la información es fácilmente adaptable a las necesidades de organizaciones de tamaño mediano y pequeño y sin esta información muchas instituciones pequeñas no podrían emprender tales programas de preservación.

El *Archivo Internet* ha venido desarrollando soluciones de código abierto para operaciones de cosecha remota que no requieren de recurso monetario, pero sí requieren de un conocimiento tecnológico aceptable. Los Archivos Nacionales de Reino Unido han llevado a cabo investigación para el mejor medio de almacenamiento, elaboraron una guía simple para archivar sitios web, y también han investigado sobre los formatos de archivo óptimos para la creación de datos. Los Archivos Nacionales de Australia han desarrollado investigaciones sobre requisitos de metadatos que son clave para manejar de forma efectiva todos los documentos de archivo, incluyendo aquellos basados en la actividad soportada en web. También han investigado sobre soluciones para registrar evidencia de documentos de archivo basados en web en sitios que cambian frecuentemente cuando raramente existen rastreadores. La Biblioteca del Congreso también ha investigado sobre metadatos específicamente para preservación (Premis es un diccionario de datos y de soporte de esquemas XML para los metadatos medulares que sustenten la preservación de materiales digitales en el largo plazo), y también realizaron investigación para desarrollar otros esquemas de metadatos Estándar para Codificación y Trasmisión de Metadatos (METS, por sus siglas en inglés), es una estructura de metadatos para metadatos de codificación, descriptivos, administrativos y estructurales

que produce los Auxiliares de Búsqueda Descriptivos de Codificación Archivística.

Desarrollar un plan de acción para la preservación del sitio web[52]

Se desarrolló un plan que incluye estrategias, protocolos, requisitos, procedimientos y resultados esperados.

[52] Véase la sección 3.0 para el Plan General de Acción para la Preservación de Sitios Web que puede ser adaptado para uso organizativo.

Preguntas de repaso

1. Nombre entre tres a seis criterios que deberían ser considerados cuando se seleccionan formatos de archivo para la producción de datos.
2. ¿Cuáles son los dos tipos de sitios web y de contenidos de sitios web y cómo difieren?
3. Nombre y describa cada uno de los tres métodos de captura de sitios web y analice cómo difieren.
4. Nombre tres grandes organizaciones que tienen buenos recursos para archivar sitios web y describa sus experiencias.
5. ¿Cuándo se deben llevar a cabo verificaciones de validación?
6. Identifique tres factores que afectan la longevidad de los medios electrónicos.
7. ¿Cuáles son los diez pasos principales en el plan general de acción para la preservación de sitios web?

Recursos adicionales**

Autor: Brown, Adrian
Título: *Archiving Websites (Archivando sitios Web)*
Fecha de publicación: 2006
Fuente/editor: Londres: Facet Publishing
Este libro es un texto amplio que ofrece guía práctica sobre programas para archivar sitios web. Se trata de una combinación de mejores prácticas, consejos prácticos y guía para establecer un programa para archivar sitios web; desde los aspectos legales y los métodos de recolección al programa de gestión y una mirada a las tendencias futuras. La obra es un recurso valioso para los profesionales de la información y gestión archivística que buscan desarrollar un plan o programa para archivar sitios web.

Autor: Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze and Sandra Payette
Título: *Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism* (*Administración de Riesgos de la Preservación de Recursos de Web. Control Virtual Remoto en el Proyecto Prism de Cornell*)
Fecha de publicación: 2002
Fuente: D-Lib Magazine 8(1)
URL: http://www.dlib.org/dlib/january02/kenney/01kenney.html

Autor: Library of Congress (Biblioteca del Congreso de EUA)
URL: http://www.digitalpreservation.gov/
La Biblioteca del Congreso de EUA tiene una variedad de excelentes programas y recursos que ofrecen información y ejemplos sobre requisitos para la preservación digital,

** Nota del traductor: Además del material que se recomienda en la versión en inglés, en español hay obras disponibles que ha elaborado la Benemérita Universidad Autónoma de Puebla.

investigación y mejores prácticas que incluyen la National Digital Stewardship Alliance, el Digital Preservation Outreach y el Education and the National Digital Information Infrastructure and Preservation Program.

Autor: National Archives of Australia
Título: *Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government (Archivando recursos de web: lineamientos para gestionar documentos de archivo de la actividad basada en web en el gobierno de la Commonwealth)*
Fecha de publicación: marzo, 2001
URL: http://www.naa.gov.au/Images/archweb_guide_tcm16-47165.pdf

Autor: Shepherd, Elizabeth y Geoffrey Yeo
Título: *Managing Records: A Handbook of Principles and Practice* (*Gestión de Documentos de Archivo: Un Manual de Principios y Prácticas*)
Fecha de publicación: 2003
Editor: Londres: Facet Publishing
Este libro es un texto amplio que delinea los principios de la gestión de documentos de archivo y su implementación práctica en organizaciones. Es exhaustivo en su cobertura respecto de los conceptos de gestión archivística. Los temas incluyen: contexto organizacional; clasificación, producción y captura de documentos de archivo, valoración, retención y disposición, acceso e implementación. El libro incluye una bibliografía amplia de recursos para la gestión archivística así como listas de estándares nacionales e internacionales sobre gestión archivística y organizaciones profesionales para gestores de documentos y archivistas.

Autor: The Internet Memory Foundation
URL: http://internetmemory.org/en/

La Fundación Memoria de Internet es una institución sin fines de lucro que apoya activamente la preservación de internet como medio para propósitos patrimoniales y culturales.

Autor: The Internet Archive
URL: http://internetmemory.org/en/
El Internet Archive desarrolló "Archive-It" (analizado en este módulo) el cual tiene archivadas cerca de 150 millones de páginas web desde 1996. También hospeda un blog sobre el desarrollo del trabajo en curso de la Máquina para Remontar el Tiempo. (Wayback Machine): http://iaWebarchiving.wordpress.com/

Autor: The National Archives UK
Título: *Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism* (*Archivado de Recursos Web: Lineamientos para Gestionar Documentos de Archivo de la Actividad basada en Web en el Gobierno de la Commonwealth*)
URL: http://www.nationalarchives.gov.uk/news/734.htm
Los Archivos Nacionales de Reino Unido son una fuente valiosa de información sobre la captura de sitios web, particularmente para capturar sitios del gobierno de Reino Unido.

Referencias bibliográficas

Blue Squirrel, *Grab-a-Site Product Page*. Disponible en: http://www.bluesquirrel.com/products/grabasite

Brown, Adrian, *Digital Preservation Guidance Note 2: Selecting Storage Media for Digital Preservation*. Agosto, 2008, Disponible en: http://www.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

Brown, Adrian, *Selecting File Formats*. Disponible en: http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf

Brown, Adrian, *Archiving Websites*, Londres: Facet Publishing, 2006.

Canadian Conservation Institute, *Electronic Media Collections Care for Small Museums and Archives*. Disponible en: http://www.cci-icc.gc.ca/caringfor-prendresoindes/articles/elecmediacare/index-eng.aspx

Digital Preservation Europe, *Database Preservation: The International Challenge and the Swiss Solutio*, octubre 2008. Disponible en: http://www.digitalpreservationeurope.eu/publications/briefs/database_preservation.pdf

European Electronic Resource Preservation and Access Network (Erpanet), *Digital Preservation Policy Tool*, 2003. Disponible en: http://www.erpanet.org/guidance/docs/ERPANET-PolicyTool.pdf

Greenwood, David J. and Morten Levin, "Reconstructing the Relationships between Universities and Society through Action Research" en: Norman K. Denzin and Yvonna S.

Lincoln, eds., *The Landscape of Qualitative Research: Theories and Issues* 2nd ed. Thousand Oaks: SAGE Publications, 2003, 131-166.

Harvey, Ross. *Preserving Digital Materials,* Munich, Germany: K. G. Saur, 2005.

Herramienta de Adobe para captura de web. Información de producto disponible en: http://ww.adobe.com/products/acrobat/

HTTrack Website. Disponible en: http://www.httrack.com/ Internet Archive, Heritrix Website. Disponible en: http://crawler.archive.org

InterPARES 3 Project, *Case Study 09 Alma Mater Society of the University of British Columbia.* Final Report.

ISO 19005-1:2005 *Document Management – Electronic document file format for long term preservation-Part 1: Use of PDF 1.4 (PDF/A-1).*"

Jones, Maggie and Neil Beagrie, *Preservation Management of Digital Materials: A Handbook.* Londres: The British Library, 2001.

Kenney, Anne R., Nancy Y. McGovern, Peter Botticelli, Richard Entlich, Carl Lagoze y Sandra Payette, "Preservation Risk Management for Web Resources. Virtual Remote Control in Cornell's Project Prism", en *D-Lib Magazine* 8(1) (2002). Disponible en: http://www.dlib.org/dlib/january02/kenney/01kenney.html

Lazinger, Susan S., *Digital Preservation and Metadata. History, Theory, Practice.* Englewood, CO: Libraries Unlimited, 2001.

Library of Congress (United States). Disponible en: http://www.digitalpreservation.gov/

Linden, Jim, Sean Martin, Richard Masters y Roderic Parker, "The Large-Scale Archival Storage of Digital Objects," en *DPC Technology Watch Series Report 04-04,* Digital Preservation Coalition, february 2005. Disponible en: http://www.dpconline.org/docs/dpctw04-03.pdf

McGovern, Nancy, Anne R. Kenney, Richard Entlich, William R. Kehoe y Ellie Buckley, "Virtual Remote Control. Building a Preservation Risk Management Toolbox for Web Resources", en: *D-Lib Magazine* 10(4), 2004. Disponible en: http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html

National Archives of Australia, *Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government,* marzo, 2001. Disponible en: http://www.naa.gov.au/Images/archweb_guide_tcm16-47165.pdf

National Archives of Australia, *Digital Recordkeeping Self-assessment Checklist,* 2004.

Pantalony, Rina Elster, *Protecting your Interests: A Legal Guide to Negotiating Website Development and Virtual Exhibition Agreements,* Ottawa, Canada, Minister of Public Works and Governments Services Canada, 1999.

Peters McLellan, Evelyn, "General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation", *InterPARES 2* Project, marzo 2007. Disponible en: http://www.interpares.org/display_file.cfm?doc=ip2_gs11_final_report_english.pdf

Prom, Christopher J. y Ellen D. Swain (2007). "From the College Democrats to the Falling Illini: Identifying, Appraising, and Capturing Student Organization Websites", en *American Archivist 70*, pp. 344-363.

Smiraglia, Richard P., *"Metadata: A Cataloger's Primer"*. Nueva York, NY: The Hawthorn Press, 2005.

Society of American Archivists. *Conference / Workshop Calendar 2009*.

Swiss Federal Archives, *SIARD Format Description*. Disponible en: http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=de