Enl@ce: Revista Venezolana de Información, Tecnología y Conocimiento
ISSN: 1690-7515
Depósito legal pp 200402ZU1624
Año 10: No. 1, Enero-Abril 2013, pp. 53-72

Cómo citar el artículo (Normas APA):
Ospina Torres, M.H. y León Luna, C.P. (2013). Una arquitectura basada en software libre para archivos web. *Enl@ce Revista Venezolana de Información, Tecnología y Conocimiento*, 10 (1), 53-72

Una arquitectura basada en software libre para archivos web

Mercy H. Ospina Torres[1]
Claudia P. León Luna[2]

Resumen

Los archivos web son sistemas de información que se han venido desarrollando desde finales de los años 90 para llevar a cabo la preservación histórica del patrimonio web como parte del patrimonio digital de la humanidad. Tales archivos han tenido que afrontar ciertos desafíos propios de los recursos web, como son el tamaño de la web y su naturaleza cambiante, las tecnologías asociadas a la web, la web superficial y la web profunda, la organización de la web en dominios, entre otros. Debido a ello, se ha hecho necesario proponer arquitecturas, técnicas, herramientas y estándares para las diferentes funcionalidades de un archivo web que permitan afrontar de manera eficaz dichos desafíos. Este trabajo tiene como objetivo establecer una arquitectura basada en software libre para un prototipo de archivo web. Para ello se hace una revisión detallada del dominio de archivo web, de sus funciones y de los enfoques usados hasta el momento para llevarlas a cabo. Se presenta un estudio comparativo entre diferentes iniciativas de preservación web a nivel mundial y se establecen los componentes para un sistema para la preservación web basada en software libre.

Palabras clave: archivo web, preservación web, software libre, arquitectura de software

Recibido: 28/2/13 Aceptado: 25/3/13

[1] Licenciada en Computación. Cursante de la Maestría en Ciencias de la Computación. Profesora de la Escuela de Computación de la Facultad de Ciencias de la Universidad Central de Venezuela.
Correo electrónico: mercy05@gmail.com

[2] Doctora en Ciencias de la Computación. Universidad Central de Venezuela. Docteur en Informatique. Université P&M Curie, Paris VI, Francia. En-Cotutela. Magister en Ciencias de la Computación. Licenciada en Computación. Investigadora acreditada al Programa de Estímulo a la Investigación, ONCTI. 2011–2013. Profesora de la Escuela de Computación, Facultad de Ciencias, Universidad Central de Venezuela. Coordinadora del Postgrado en Ciencias de la Computación, UCV, 2012-2014.
Correo electrónico: claudia.leon@gmail.com

An Architecture Based on Software Free for Web Files

Abstract

The Web Files are information systems that they have come developing from ends of the 90s to carry out the historical preservation of the web patrimony as part of the digital heritage of the humanity. Such files have had to confront certain proper challenges of the web resources, as there are the size of the Web and his changeable nature, the technologies associated with the web, the superficial web and the deep web, the organization of the Web in domains, between others. Due to it, it has become necessary to propose architectures, skills, hardware and standards for the different functionalities of a Web File that allow to confront in an effective way the above mentioned challenges. This work takes as a target to establish an architecture based on software free for a prototype of Web File. For it there does a detailed review of the mastery of Web File, of his functions and of the approaches used until now to carry out them. A comparative study appears between different initiatives of web preservation on a global scale and the components are established for a system for the web preservation based on free software.

Key words: Web File, Web Preservation, Free Software, Architecture of Software

Introducción

La información pública disponible en la web hoy en día es más grande que la información difundida en cualquier otro medio de comunicación. Sin embargo, su permanencia no está garantizada debido a la alteración o eliminación de los documentos web o de sus contenidos (Masanés, 2006). Como se muestra en el mapa conceptual de la **Figura 1**, un documento web se define como un documento basado en el lenguaje de marcas HTML (HiperText Markup Language) también llamado página web, así como los demás archivos asociados en su composición, tales como imágenes, videos, hojas de estilo, entre otros. Dicho documento puede ser localizado a través de un URL[3] y en general pertenece a un sitio web, además, puede referenciar a otros documentos web.

Desde finales de los 1990 se han desarrollado sistemas de información cuyo objetivo es preservar conjuntos seleccionados de sitios web, denominados archivos web, que deben su nombre ya que se les puede considerar una especialización de los archivos históricos[4] comúnmente conocidos por su rol en la preservación documental y que también abarcan los archivos digitales, donde los

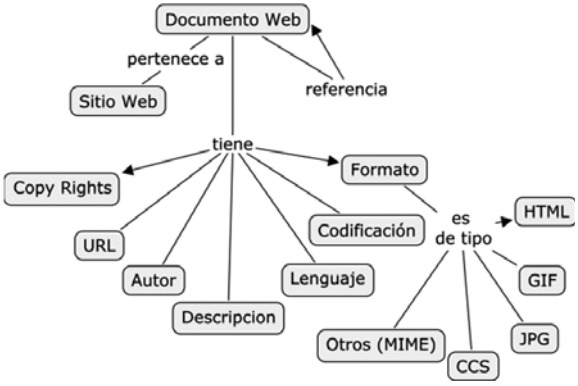
[3] Localizador uniforme de recursos (Uniform Resource Locator) estándar propuesto por Tim Berners-Lee para permitir establecer hiperenlaces en los documentos web.

[4] Según el Consejo Internacional de Archivos, un conjunto de documentos, sea cual sea su forma o soporte material, producidos por personas u organismos públicos o privados, que tienen un valor informativo, histórico y cultural, que son conservados a perpetuidad en condiciones que garanticen su integridad y transmisión a generaciones futuras.

documentos preservados son únicamente de origen digital.

Figura 1

Mapa conceptual de documentos web

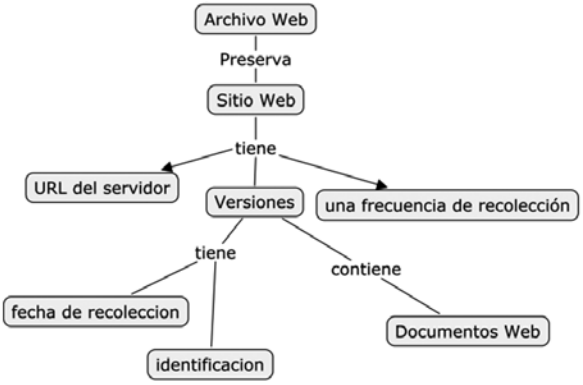


La preservación realizada por los archivos web se logra mediante la replicación y/o migración de los documentos que conforman los sitios web, desde su formato original a otra representación. Los sitios replicados son mantenidos completos, es decir, acompañados de los archivos, imágenes, gráficas y el aspecto visual. Son almacenados en servidores de preservación en un ambiente seguro (Masanés, 2006).

El mapa conceptual de la **Figura 2** resume la definición de archivos web, donde las versiones de un sitio web representan su estado en un instante de tiempo dado y reflejan el hecho de que el contenido y la estructura de los sitios web cambian con el tiempo.

Figura 2

Mapa conceptual de preservación de sitios web en archivos web



Los archivos web que se han desarrollado desde finales de los 1990 (también conocidos como iniciativas de preservación web) han propuesto diversos enfoques para sus funciones y arquitectura (Ball, 2010; Hwang, Kim, & Singh, 2007; Library of Congress of USA, 2002; Masanés, 2006; Lyman, 2002; Strodl, Becker, Neumayer, y Rauber, 2007), motivo por el cual el Consorcio Internacional de Preservación de Internet (IIPC por sus siglas en inglés), actual ente regulador para la preservación web (IIPC, 2011), ha definido una arquitectura funcional para los archivos web.

En el presente trabajo se propone una arquitectura para archivos web basada en la arquitectura funcional del IIPC como parte de un marco de trabajo que pueda ser utilizado por entes gubernamentales o institucionales que requieran crear

un archivo web para la preservación web local. Este artículo está estructurado en las siguientes secciones: en la sección 2 se describe la arquitectura definida por el IIPC. En las secciones 3 y 4 se hace una descripción de los principales métodos de adquisición y de almacenamiento que pueden ser usados por los archivos web. En la sección 5 se presenta un estudio comparativo de distintas iniciativas de preservación web a nivel mundial en cuanto al uso de los métodos descritos en las secciones anteriores, así como las herramientas y estándares usados. El estudio tiene como objetivo facilitar la toma de decisiones para la implementación de la arquitectura a proponer. Por último, en la sección 6 se describe la arquitectura de archivo web propuesta.

Arquitectura funcional de la IIPC

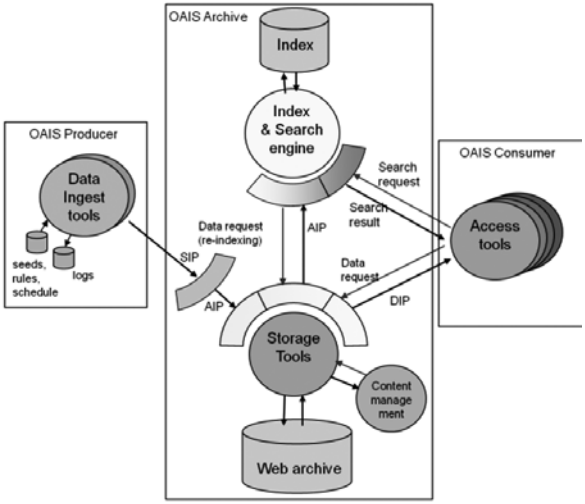
Las tareas que deben ser llevadas a cabo por los archivos web para cumplir su objetivo de preservación se describen en Masanés (2006) y son las siguientes:

- Selección de las páginas o sitios web a resguardar
- Adquisición regular del contenido de dichas páginas
- Almacenamiento e indexación de las páginas resguardadas
- Recuperación o consultas sobre la información resguardada

El IIPC propone una arquitectura funcional para archivos web que se muestra en la **Figura 3**, donde se establecen los elementos necesarios para llevar a cabo estas tareas. Esta arquitectura a su vez se basa en modelo de referencia Open Archive Information System (OAIS), el cual proporciona un marco de alto nivel para el desarrollo de archivos digitales. Fue propuesto por el Comité Consultivo de Sistema de Datos del Espacio (Consultative Committee for Space Data Systems, CCSDS) (CCSDS, 2002) y aprobado por la Organización Internacional de Normalización como norma ISO14721:2003.

Figura 3

Arquitectura funcional IIPC basada en el modelo OAIS (Masanés, 2006)

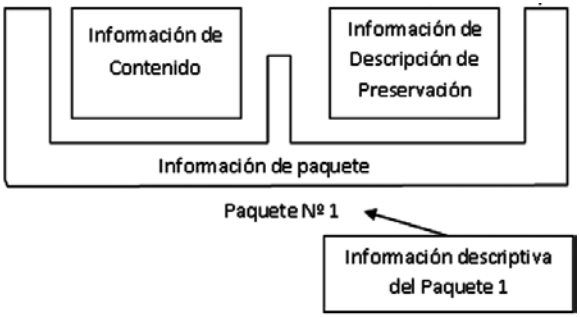


Para esta arquitectura, el modelo OAIS define dos roles que interactúan con el archivo, los cuales son**: el productor** que son las personas o sistemas de clientes que proporcionan la información a ser conservada, y **el consumidor,** que

son las personas o los sistemas de clientes que interactúan con el archivo para encontrar y adquirir información de interés conservada. A su vez, la información que se intercambia entre los roles y se almacena en el archivo se denomina paquete de información (IP por sus siglas en inglés, *Information Package*) cuya estructura se muestra en la **Figura 4**.

Figura 4

Paquete de información (CCSDS, 2002)



Cada IP está compuesto por la información de contenido, que es el conjunto de información original de preservación, y la información de descripción de preservación, que permite comprender y contextualizar la información de contenido. A su vez, la información de descripción de la preservación está formada por:

- Información de referencia: para la identificación de la información de contenido
- Información de origen que indica la fuente de la información de contenido
- Información de contexto para relacionar la información de contenido con otros IP

A medida que estos IP son adquiridos y procesados por el archivo pueden sufrir transformaciones, por lo que el modelo define tres tipos de IP. El IP recibido del productor o Submit IP (SIP), el IP almacenado en el Archivo o Archive IP (AIP) y el IP entregado al cliente o Delivery IP (DIP). En la figura 3 se muestran estos tipos de IP y su interacción con los componentes de la arquitectura.

Los mencionados componentes a su vez dan soporte a las tareas de los archivos web nombradas al inicio de esta sección de la siguiente manera: las herramientas de ingreso de datos (*Data ingest tools*) se encargan de la adquisición regular de los sitios web seleccionados y sus documentos web asociados. El proceso de selección, aunque forma parte de las políticas y lineamientos de cada archivo, es apoyado por las herramientas de ingreso a través de un repositorio de los sitios a preservar y la frecuencia de cambio para la preservación de nuevas versiones. Para la indización de los documentos se hace uso de la tecnología de motores de búsqueda e indización (*Index & Search engine*), desarrolladas por los buscadores de internet como Google[5], Altavisita[6] y Yahoo[7]. El almacenamiento es llevado a cabo por las herramientas de almacenamiento (Storage Tools) y los manejadores de contenido (*content management*). Por úl-

[5] www.google.com
[6] www.altavista.com
[7] www.yahoo.com

timo, el acceso a los datos se hace a través de las herramientas de acceso (*Access tools*).

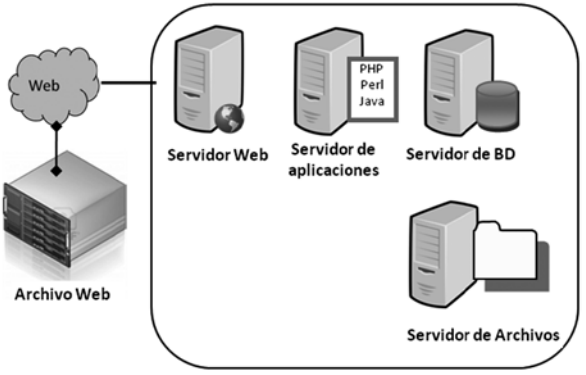
En la actualidad existe un conjunto de herramientas y estándares de código abierto disponibles para llevar a cabo algunas de las tareas antes descritas. La selección de estas herramientas y estándares depende de las necesidades y tipos de archivo.

Métodos de adquisición de datos

Adquisición del lado del cliente

La adquisición del lado del cliente se lleva a cabo usando rastreadores o copiadores de sitios web, herramientas derivadas y adaptadas de las tecnologías de los motores de búsqueda. Este método es muy usado debido a su simplicidad y escalabilidad, así como por su adaptación al entorno cliente-servidor. Este método adopta la misma posición de los usuarios de la web e imita su forma de interacción con los servidores. Los rastreadores comienzan desde las páginas web semilla, es decir los puntos de entrada de los sitios web a preservar, y realizan una solicitud HTTP (*HiperText Transfer Protocol*) al servidor web y éste a otros servidores. Dependiendo del caso, tal como se presenta en la **Figura 5**, la respuesta recibida es analizada, se extraen los enlaces o vínculos y se buscan los documentos enlazados. Este proceso se repite con los documentos traídos hasta que no haya vínculos por explorar dentro del alcance definido. Esto es debido a que el protocolo HTTP no proporciona un comando que devuelva la lista completa de los documentos disponibles en el servidor, como por ejemplo, el protocolo FTP (*File Transfer Protocol*). Cada página tiene, por lo tanto, que ser "descubierta" por extracción de enlaces desde otras páginas.

Figura 5
Adquisición del lado del cliente



En este enfoque es necesario adaptar la tecnología de rastreo de manera que cumpla con los objetivos de preservación. El primer cambio es que se deben tratar de buscar todos los archivos “sin importar” su formato, incluyendo aquellos formatos que puedan ser muy pesados y que no serían tomados en cuenta por los rastreadores tradicionales, para permitir archivar una versión completa de los sitios. El segundo cambio se refiere a las normas de cortesía de los rastreadores que buscan evitar la sobrecarga de los servidores rastreados lo que implica que la captura de un sitio pueda durar minutos o días en algunos casos. Esto puede traer como consecuencia que las páginas ya copiadas en el archivo web pudieran haber sido cambiadas en el servidor durante el rastreo,

lo que puede generar una inconsistencia temporal del archivo web.

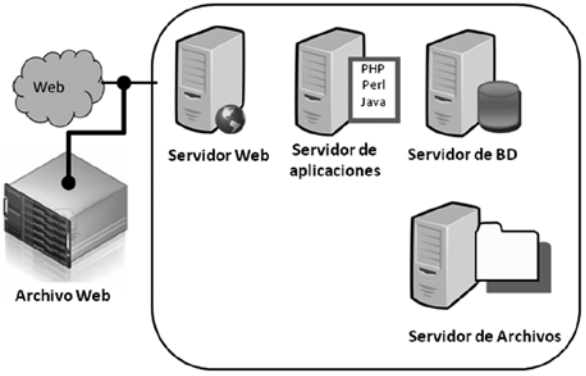
La mayoría de los problemas relacionados con este método ocurren durante la extracción de enlaces debido a URLs mal formados o que usan parámetros complejos. Otros pueden ser causados por redirecciones, autorizaciones, respuestas lentas y respuestas no válidas del servidor, entre otras. Una de las limitaciones de este método es que solamente rastrea la web superficial[8]

Archivado de transacciones

El archivado de transacciones, propuesto por Fitch (2003), consiste en la captura y archivado de "*todas las respuestas materialmente distintas producidas por un sitio web, independientemente de su tipo de contenido y de cómo se producen*", tal como se muestra en la **Figura 6**. En este caso se almacenan los pares solicitud/respuesta únicos, y por lo tanto se crea un archivo completo de todos los contenidos vistos para un sitio específico, de manera que las peticiones con apenas leves diferencias son consideradas como únicas.

Este tipo de archivo web resulta útil para rastrear y registrar todas las instanciaciones posibles de contenido. Entonces, el contenido nunca visto no será archivado pero el contenido web oculto, siempre y cuando se acceda, será almacenado, lo que presenta una ventaja en ese respecto. Este método además tiene la ventaja de permitir la grabación de exactamente lo que se ve y cuándo, lo cual puede ser necesario en corporaciones e instituciones motivadas por la responsabilidad legal. La principal limitación de este método es el hecho de que tiene que aplicarse con el consentimiento y la colaboración del propietario del servidor. Por tanto, es indicado principalmente para el archivado web en intranet de organizaciones.

Figura 6
Adquisición de transacciones



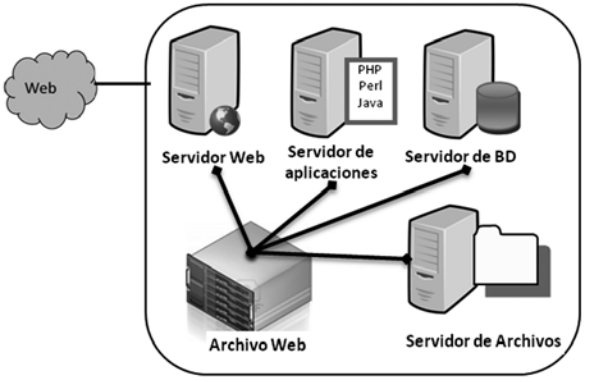
Archivado del lado del servidor

El archivado del lado del servidor consiste en copiar los archivos directamente desde el servidor, sin usar la interfaz HTTP en absoluto, tal como se muestra en la **Figura 7**. Este método, como el anterior, únicamente se puede utilizar con la colaboración de los propietarios del sitio, y plan-

[8] La web superficial son aquellas páginas alcanzables por los rastreadores, a diferencia de la web profunda u oculta, que solo se puede acceder a través de formularios o que están protegidas por contraseña.

tea algunas dificultades para hacer que el contenido copiado sea utilizable. Esto es así pues supone la posibilidad de reproducir completamente la arquitectura y el contenido del servidor, por lo tanto requiere de la participación "activa" del administrador del sitio, pues cualquier cambio en el servidor original puede dejar el archivo inoperable.

Figura 7
Adquisición del lado del servidor



Aquí el método implica la reducción de la dependencia en la base de datos y la ejecución de scripts del lado del servidor y, aunque con su uso se puede obtener tanto la web superficial como la oculta, mantener infraestructuras para cada sitio web que se desee preservar representa un alto costo. Por ello, nada más es indicado para el archivado de web interno en organizaciones.

Métodos de almacenamiento

Hacer una copia de un sitio web es una tarea compleja la cual implica volver a crear un sistema de información que será accesible para los usuarios. Además, en un archivo web no se almacena una, sino varias copias del mismo sitio web, cada una considerada como una versión asociada a una variable temporal. Lo ideal sería que cada versión del sitio web dentro del archivo fuera isomórfica al original en el momento de la captura (la misma estructura jerárquica, mismos nombres de archivos, mismos mecanismos de enlace, mismo formato), pero por razones prácticas ello casi nunca sucede. En algunos casos, tal como se describe en la sección 3, para ser efectiva, la adquisición de sitios induce a una transformación de los documentos a preservar. Esta es la razón por la cual los responsables de los archivos web han adoptado diferentes estrategias de almacenamiento las cuales pueden afectar el direccionamiento, los mecanismos de enlace y los formatos, así como el objeto mismo de su representación.

Archivado por el sistema de archivos local

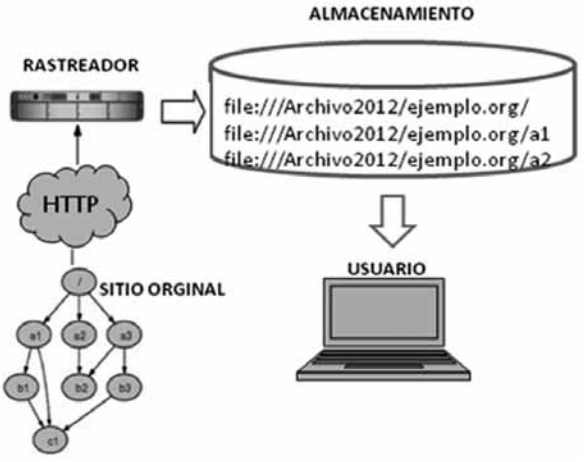
En esta estrategia, mostrada en la **Figura 8**, se copian localmente los documentos rastreados desde el sitio web y se usa la especificación URI[9] (Berners-Lee, 1994) del sistema de archivos (SA) local que usa el prefijo "*file*" para accederlos. Por ejemplo, la dirección HTTP://www.example.

[9] Identificador uniforme de recursos (Uniform Resource Identifier). Define un espacio de nombres para objetos en Internet

org/example.HTML, en el SA local pasaría a ser file:///Users/archive2005/example.org/example. HTML.

Figura 8

Archivado usando el SA local (Masanés, 2006)



El procedimiento permite el uso del SA para navegar a través del material archivado y, siempre y cuando los enlaces en los documentos sean relativos, la navegación en el archivo será la misma que en el sitio original. Será diferenciable solamente en la barra de direcciones del navegador cuando se mira en el prefijo de URI (en este caso "file" en lugar de "http"). El principal beneficio de este enfoque es la correspondencia de la estructura del sitio web original y su bajo costo en implementación y uso, debido a que los navegadores web pueden mostrar las páginas web directamente del SA local.

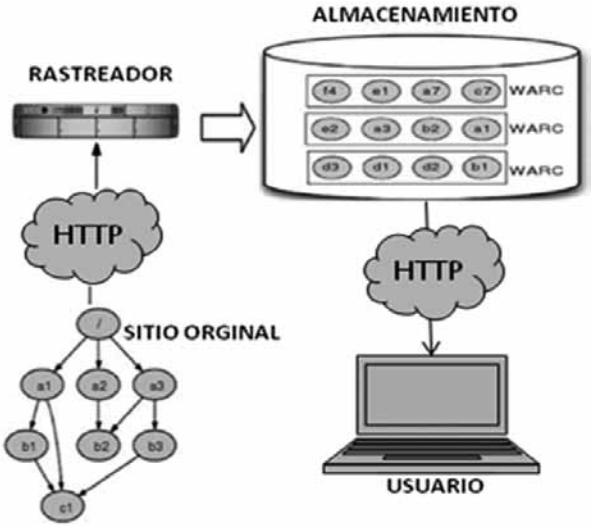
Uno de los problemas de este tipo de archivos es que para poder usar el SA local es necesario efectuar transformaciones de los documentos originales para garantizar la navegación, tales como son los cambios de los enlaces absolutos a relativos y los cambios de nombres de los objetos, por lo que no se respetaría la fidelidad estricta a la estructura original. El otro problema importante viene dado por la estructura jerárquica del SA, ya que se deben mantener y manejar las distintas versiones de cada sitio web, por lo que realizar una correspondencia de esta organización con la estructura jerárquica del SA no se puede llevar a cabo directamente sin realizar cambios. Otra limitación de este enfoque viene dada por la gran cantidad de documentos que un archivo web debe manejar. Es así como un archivo web puede contener miles de millones de documentos, cifra que llega al límite de la capacidad de los actuales "sistemas de archivos" e incluso aun cuando se pueda manejar esta cantidad de archivos, el rendimiento puede verse afectado.

Archivado con servidor web

Se basa en archivar la respuesta a diferencia del primero que archiva cada documento por separado. Las respuestas del servidor original se almacenan sin cambios en un contenedor de archivos WARC[10] que permite servirlos más tarde a los usuarios del archivo con un servidor HTTP, tal como se muestra en la **Figura 9**.

[10] Formato propuesto por Internet Archive, estandarizado por IIPC y presentado por ISO.

Figura 9
Archivado con servidor web



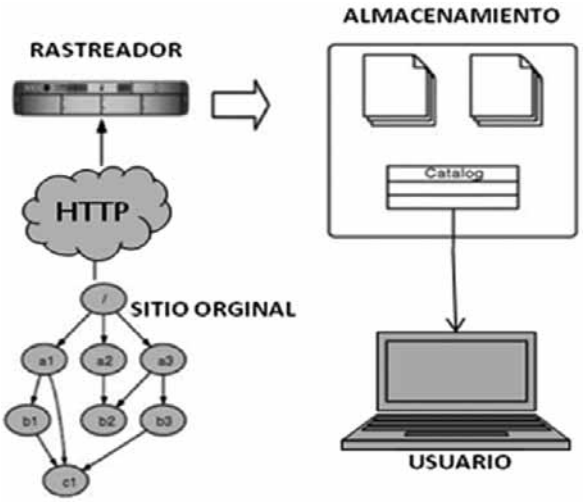
Un archivo WARC registra una secuencia de documentos web asociados con una actividad de recolección. Cada página está precedida por un encabezado que describe brevemente el contenido de cosecha y su longitud. En ISO (2009) se define a WARC como una norma ISO. Una de las características de este método es que se conservan el esquema de nombres original (incluyendo los parámetros de las páginas dinámicas) y permite la navegación en el sitio tal cual como ha sido rastreado. El usuario del archivo puede recorrer de nuevo todos los caminos seguidos por el rastreador. Su principal ventaja es la posibilidad de superar la limitación de los SA en términos de tamaño y espacio de nombres. La desventaja es que es imposible el acceso directo a los archivos almacenados, por lo que se requiere un índice para los contenedores y un servidor web que muestre el contenido de dichos contenedores. Estas dos capas requieren la ejecución de un ambiente de acceso el cual puede ser difícil de configurar y de mantener.

Archivado no web

En este enfoque, el cual se ilustra en la **Figura 10**, los documentos que están en la web son extraídos del contexto de hipertexto y reorganizados en un estilo diferente en términos de lógica de acceso y / o formato. Este puede ser el caso cuando un conjunto de documentos sacados de la web es reorganizado a partir de una lógica de acceso basada en enlaces, a una lógica basada en catálogo. También cuando una página o un sitio web completo se transforma en formato PDF. En este caso, el documento está prácticamente impreso, lo que implica una representación congelada y una organización similar a la página de papel. Incluso los enlaces puede funcionar con un esquema interno y de propiedad de nombres.

El enfoque tiene sentido sobre todo para los objetos que han sido originalmente creados y organizados de forma independiente de la web. Es el caso, por ejemplo, de grandes colecciones de libros digitalizados, documentos, música, vídeos disponibles en la web, donde la organización original no era hipertextual, sino basada en catálogo.

Figura 10
Archivado no web



Estudio de iniciativas de preservación web

Para poder tomar decisiones en cuanto a las herramientas, las técnicas, los estándares y la infraestructura a ser usadas para el desarrollo de un archivo web se recopilaron datos de 52 iniciativas de preservación web oficiales conocidas, las cuales permitieron realizar un estudio comparativo entre los aspectos mencionados. Esta recopilación se basó en un estudio previo realizado por el Archivo web Portugués (Gomes y Costa, 2011) donde se publican datos de 42 iniciativas a nivel mundial los cuales se actualizaron y completaron revisando las páginas web de las diferentes iniciativas estudiadas.

Las variables en estudio son:

- País de origen
- Año de creación
- Dominios web que preserva
- Métodos y herramientas de adquisición de datos
- Métodos y estándares de almacenamiento
- Cantidad de almacenamiento usado
- Tipos de búsqueda
- Organizaciones que promueven el archivo

Del análisis de estas variables se observó que:

El 59% de las iniciativas se hospedan en Europa, el 23% en Norteamérica y las restantes en Asia y Oceanía. Se destaca que no existían iniciativas web en Latinoamérica promovidas por los gobiernos para el momento de este estudio. En la **Tabla 1** se muestra el porcentaje de uso de los métodos y las herramientas de adquisición, donde se observa que el método de adquisición más usado es el del lado del cliente, descrito en la sección 3.1. Se observa además que la herramienta más utilizada en la adquisición es el rastreador Heritrix (Mohr, Stack, Ranitovic y Kimpton, 2004), con un 67%, el cual será descrito en la sección 6. Entre las iniciativas que la usan se supo que algunas utilizaron copiadores de sitios web como Httrack (HTTrack, 2011) o Wget (GNU Operating System, 2012). Sin embargo luego migraron a Heritrix.

Tabla 1

Resumen de métodos y herramientas de adquisición usadas por iniciativas de preservación web a nivel mundial

Método de adquisición		**%**	**Herramientas**		**%**
Del lado del cliente	47	90%	Hanzo Crawler	4	8%
			Heritrix	35	67%
			PhagoSite	2	4%
			Propio	6	12%
No conocido	5	10%		5	10%
Total	**52**			**52**	

El porcentaje de uso de los métodos y estándares de almacenamiento se muestran en la **Tabla 2**. Allí se observa que el almacenamiento con servidor web presenta un uso del 73% (38 iniciativas), y de las 14 restantes se puede afirmar que 4 iniciativas (8% del total) usan el SA local. Se desconoce el método utilizado en las otras diez.

Tabla 2

Resumen de métodos y estándares de almacenamiento usados por iniciativas de preservación web a nivel mundial

Método de Almacenamiento		**%**	**Estándar**		**%**
Con servidor Web	38	73%	WARC/ARC	36	69%
			DIFF	2	4%
SA local	4	8%	file://	4	8%
No conocido	10	19%		10	19%
Total	**52**			**52**	

De las iniciativas que usan almacenamiento con servidor web, 36 usan los estándares ARC/WARC y dos usan un formato propio denominado DIFF (Ben Saad, Gançarski y Pehlivan, 2009). En cuanto a la selección de las páginas a archivar, 47 (90%) recolecta páginas web locales asociadas al dominio de su país o región clasificadas en temas de importancia para cada gobierno, tales como cultura, educación, tecnología, investigación, noticias, historia, entre otros. Tres de los cinco restantes: Internet Archive (Internet Archive, 2012), European Archive (Internet Memory Foundation, 2010) y Hanzo Archive (Hanzo Archives, 2009) recolectan documentos web de múltiples dominios, ofreciendo sus servicios de recolección a sus miembros. Los dos restantes son archivos comerciales que ofrecen sus servicios principalmente a empresas.

De las iniciativas locales, 30 (57%) fueron desarrolladas bajo el auspicio de la Biblioteca o los Archivos Nacionales del país o de alguna universidad. Con respecto al volumen de información manejada, la iniciativa con mayor volumen es el Internet Archive, el cual posee hasta la fecha alrededor de 2,5 Petabyes de información almacenada. Va seguido por la Universidad de Texas con 1,1 Petabyes y por el European Archive con 0,5 Petabytes. Excluyendo estas tres iniciativas, el promedio de espacio de almacenamiento requerido es de aproximadamente 40 terabytes.

En cuanto a la tecnología de indización y búsqueda, la cual se muestra en la **Tabla 3**, se observa que 36 utilizan motores de búsqueda e indización basados en la librería Lucene de Apache (Apache, 2012), entre estos está el NutchWAX (Internet Archive, 2008) que es el motor de búsqueda desarrollado y usado por Internet Archive con 54%. Hanzo Engine con 8%, Apache Solr con 4% y Lucene con 4%. Las demás iniciativas usan motores propios o no dan información.

Tabla 3

Resumen de indizadores usados por iniciativas de preservación web a nivel mundial

Tecnología de Indización y búsqueda		**%**
NutchWAX	28	54%
Solr	2	4%
Lucene	2	4%
Hanzo Engine	4	8%
Propio	7	13%
Sin información	9	17%
Total	**52**	

En resumen, el estudio realizado muestra que el método de adquisición más usado es la adquisición del lado del cliente utilizando como principal herramienta el rastreador Heritrix. Como método de almacenamiento se usa principalmente el almacenamiento con servidor web con contenedores ARC o WARC. Para la indización y búsqueda se utiliza algún motor basado en la Librería Apache Lucene. Sobre la base de estos resultados se realizó la selección de herramientas y estándares para la definición de la arquitectura de archivo web propuesta.

Arquitectura de archivo web propuesta

La importancia de la preservación web ha sido ampliamente estudiada, y organismos como la UNESCO (UNESCO, 2003) y el IIPC han realizado esfuerzos para fomentar el desarrollo y uso de archivos web. Debido al tamaño de la web y a la naturaleza histórica de los archivos web, la mayoría de las iniciativas estudiadas almacenan una porción de la web relacionada con el dominio local. Sin embargo, aún existen muchos países que no cuentan con iniciativas oficiales de preservación web conocida, tal como es el caso de Venezuela. Únicamente se conoció de la investigación realizada por el Centro de Computación Paralela y Distribuida (CCPD) de la Universidad Central de Venezuela, donde se describen los lineamientos para la construcción de un archivo histórico de la información digital producida en Venezuela. En dicha investigación se proponen técnicas y lineamientos para la adquisición de documentos (Sanoja, León y Torres, 2010). El proyecto es una primera experimentación, con técnicas semiautomáticas de selección y adquisición que sirven de referencia al presente trabajo.

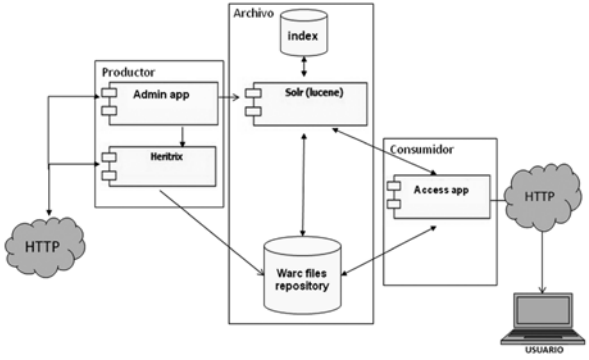
La arquitectura propuesta representa un archivo web con las siguientes características:

- Las páginas a recolectar serán de dominio local en Venezuela a nivel de sitios web.
- La solución presentada estará basada en tecnologías de software libre de manera de dar cumplimiento al decreto 3390 publicado en Gaceta Oficial de la República Bolivariana de Venezuela (2004) y en herramientas y estándares internacionales que han sido utilizados con éxito por otras iniciativas. En este sentido:
  - El método de adquisición a usar será del lado del cliente con alguna de las versiones del rastreador Heritrix.
  - El método de almacenamiento será almacenamiento con servidor web con contenedor de archivos WARC, por ser un estándar definido y aprobado por ISO.
- Los componentes asociados a la administración del archivo y al acceso serán desarrollados localmente usando tecnología de software libre.

En la **Figura 11** se muestra la arquitectura propuesta para un prototipo de archivo web usando notación UML.

Figura 11

Modelo de arquitectura de Archivo Web propuesta



Descripción de los componentes

Heritrix

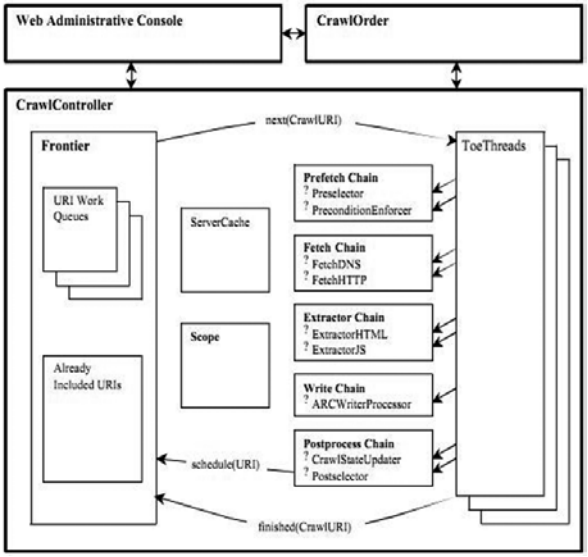
Heritrix es un rastreador (*crawler*) de documentos web a través de Internet donde, dada una URL semilla o punto de entrada de un sitio web y una configuración de rastreo, se obtienen las páginas y demás documentos web asociados a dicho sitio. Tal como se describió en la sección 3.1. su licencia es de open-source y está escrito completamente en Java. Su interfaz de configuración es accesible usando un navegador web y también puede ser lanzando desde línea de comandos usando el API Rest (Cowan, 2005). Fue desarrollado por Internet Archive a principios de 2003 con el propósito específico de rastrear para archivado de sitios web. Al ser de código abierto, ha fomentado la colaboración y el desarrollo con entes similares que necesiten servicio de rastreo. La ejecución de un rastreo sigue los siguientes pasos:

1. Elegir un URI de entre todas las programadas
2. Buscar el URI usando el protocolo HTTP
3. Analizar los archivos recibidos para extraer nuevos enlaces y/o archivar los resultados
4. Seleccionar los URI descubiertos que sean de interés y sumarlos a los ya programados
5. Terminar el procesamiento de la URI actual y repetir el proceso con las demás URIs programadas

En la **Figura 12** se presenta la arquitectura de Heritrix (Mohr, Stack, Ranitovic y Kimpton, 2004).

Figura 12

Arquitectura de Heritrix (Mohr, Stack, Ranitovic y Kimpton, 2004)



Los tres componentes más importantes son:

1. El Alcance (*Scope*): determina si cierta URI está fuera o dentro de las reglas de rastreo. El alcance incluye las semillas URI que se usan para iniciar el rastreo. El alcance también interviene en la selección de URIs mencionadas en el paso 4 del proceso de rastreo.
2. La Frontera (*Frontier*): es el responsable de seleccionar el siguiente URI a ser procesado, además de llevar un registro de la URI cosechadas y otro de las URI que ya han sido procesadas.

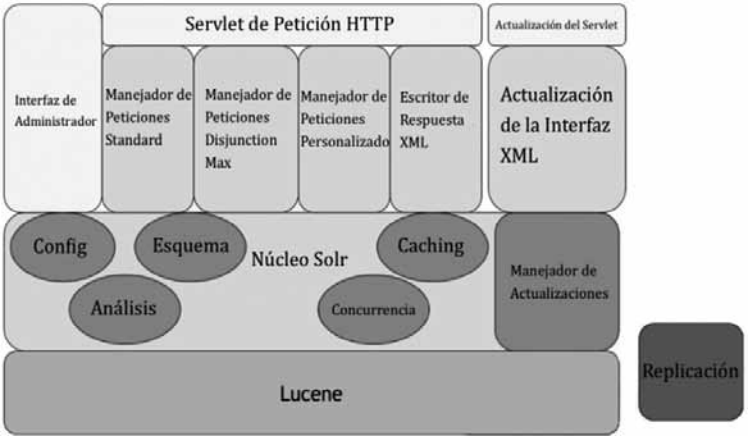
3. La cadena de procesamiento (*ProcessorChains*): incluye procesadores modulares que realizan tareas específicas en cada URI. Esto incluye la búsqueda del URI, análisis de los resultados devueltos y pase de URIs descubiertas a la frontera.

A medida que se va realizando el rastreo, los archivos se almacenan en un archivo contenedor cuyo formato se selecciona previamente y el cual puede ser ARC(Jack y Binns, 2012) o WARC.

Apache Solr

Solr es una plataforma de búsqueda de código abierto basada en el proyecto Apache Lucene, también escrita en Java, que se ejecuta como un servlet en Tomcat y que usa la librería Java de Lucene (Apache, 2012) como su núcleo de búsqueda (Smiley y Pugh, 2011). En la **Figura 13** se ilustra la arquitectura completa de Solr.

Figura 13
Arquitectura de Solr (Apache Solr, 2012)



Una de sus principales características es que permite las búsquedas de texto completo, provee búsquedas dinámicas y distribuidas, la integración con bases de datos y la replicación de índices. Para las consultas, posee una interfaz HTTP con la posibilidad de configurar las respuestas en formatos como XML/XSLT, JSON, Python, Ruby, PHP, Velocity, Binary, entre otras y hace uso del analizador de consultas DisMax (Disjunction Max) para obtener resultados de alta relevancia de consultas introducidas por el usuario.

Otras características que posee son que maneja campos estadísticos numéricos como mínimo, máximo, promedio (avg) y desviación estándar. Combina consultas derivadas de diferentes sintaxis, sugiere de manera automática opciones para completar las consultas de los usuarios, permite la configuración de los resultados de mayor puntuación para una consulta y posee optimización en cuanto al desempeño.

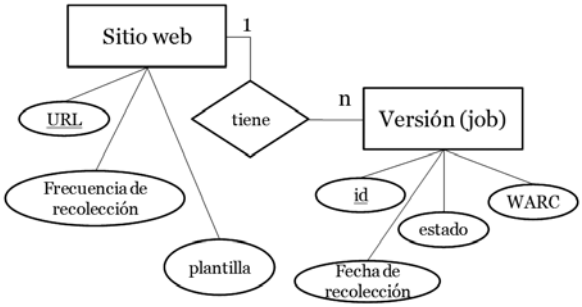
Admin app

Debido a que el rastreador Heritrix efectúa rastreos, una vez estos sean configurados y ejecutados manualmente, se propone una aplicación administrativa que tiene como objetivo automatizar el proceso de rastreo. Para ello almacena los URIs semillas a rastrear y planifica los rastreos de las versiones dada una frecuencia de rastreo de dichas semillas usando la interfaz REST de Heritrix para crear los nuevos trabajos de rastreo de manera automática y, una vez completados con éxito, notificará a Solr la culminación del trabajo para su indexación. Esta aplicación debe contar con un repositorio donde se almacenarán los URL de los sitios web a rastrear, así como su frecuencia de rastreo, además de la información de cada uno de los trabajos de rastreo efectuados sobre cada sitio web. El modelo de datos de este repositorio se muestra en la **Figura 14**.

En este modelo el atributo "plantilla" es un Archivo que puede usar Heritrix como base para configurar los valores del rastreo para un trabajo (*job*), el atributo "estado" permite saber si el trabajo está en proceso, falló, fue culminado o fue indexado, y el atributo "WARC" permite conocer la ubicación del Archivo WARC asociado al job.

Figura 14

Modelo ER del repositorio de la aplicación Admin App



La frecuencia de recolección se puede actualizar a medida que se realicen recolecciones y se detecte si estas son diferentes a la última versión almacenada, de manera que se pierda la menor cantidad de cambios en las páginas y evitar además el almacenamiento de versiones redundantes.

Acess app

Se trata de una aplicación web que permitirá al usuario final conocer los sitios web que se preservan en el archivo web y acceder a sus diferentes versiones, realizando una consulta bien sea usando el URL o por palabras clave. Para ello debe interactuar con el motor de indexación y búsqueda solr, componente que recibe la consulta y entrega la ubicación de los archivos WARC que cumplen con ésta. Estos archivos deben ser desplegados en el navegador para poder visualizar las páginas web que contienen. Igualmente, sus enlaces internos deben acceder a elementos dentro del WARC y no en la web.

Repositorio de archivos WARC

Es el espacio de almacenamiento donde se localizan los archivos WARC pertenecientes a los rastreos efectuados por sitio web y que son indexados para su acceso. Los archivos WARC usan un formato que permite almacenar, describir y guardar recursos de la web junto con los sucesivos cambios que éstos puedan experimentar, a lo largo de su exposición, en un solo archivo contenedor. Además puede controlar la información de protocolos de la capa de aplicación como HTTP, DNS y FTP y comprimir datos manteniendo su integridad (ISO, 2009). Para acceder al contenido de estos contenedores es necesario el uso de librerías o herramientas especializadas como los Wacrtools (IIPC, 2012) los cuales aún están en proyecto, haciendo complejo el desarrollo de las aplicaciones de acceso.

Conclusiones

Como respuesta a la ausencia de iniciativas de preservación web gubernamentales o privadas en Venezuela y Latinoamérica, en este trabajo se propone una arquitectura para un archivo web como parte de un marco de referencia para la creación de archivos web que utilice software libre y estándares de facto y que pueda ser asimilado por instituciones venezolanas y latinoamericanas. Debido a los múltiples enfoques existentes para el desarrollo de archivos web, se realizó un estudio de los principales métodos, estándares y herramientas usados en el área de preservación web a nivel mundial. Se asociaron los métodos con las herramientas y estándares y se determinó su porcentaje de uso por las iniciativas de preservación web estudiadas. Los resultados de este estudio permitieron realizar una escogencia de los métodos, estándares y herramientas a usar en la arquitectura propuesta, dando prioridad al uso de software libre con posibilidad de acceso al código. Actualmente está en desarrollo un prototipo funcional cuyo propósito es implementar los componentes propuestos, probar la integración e interacción con las herramientas seleccionadas y realizar su calibración para un rendimiento adecuado bajo constantes cambios de los sitios web a preservar.

Bibliografía

Apache (2012). Lucene Apache. Recuperado de http://lucene.apache.org/

Apache Solr (2012). *Solr*. Recuperado de http://people.apache.org/~sgoeschl/presentations/solr/index.html

Ball, A. (2010). Web archiving. Edimburg, UK.: Digital Curation Centre.

Ben Saad, M., Gançarski, S., y Pehlivan, Z. (2009). Archiving web pages based on visual analysis and DIFF. *25ème Journées des Bases de Données Avancées (BDA)*. Namur, Begium.

Berners-Lee, T. (1994). Universal Resource Identifiers in WWW. Recuperado de http://www.ietf.org/rfc/rfc1630.txt

CCSDS. (2002). Reference model for an open archival information system. Washington, Blue Book

Cowan, J. (2005). RESTful Web Services. Recuperado de http://home.ccil.org/~cowan/restws.pdf

Decreto 3390. (2004). *Gaceta oficial Nº 38.095 de la Republica Bolívariana de Venezuela*. de fecha 28 de Diciembre de 2004, Asamblea Nacional, Caracas, Venezuela. Fitch, K. (2003). Web site archiving: An approach to recording every materially. *AusWeb 2003: The Ninth Australian World Wide Web Conference*. Sanctuary Cove, Australia.

GNU Operating System. (2012). *GNU Wget*. Recuperado de http://www.gnu.org/s/wget/

Gomes, D., Miranda, J. y Costa, M. (2011). A survey on web archiving initiatives. En *Research and Advanced Technology for Digital Libraries*. (pp. 408-420). Springer Berlin Heidelberg

Gomes, D., Miranda, J. y Costa, M. (2010). A survey on web archiving initiatives. Portugal.

Hanzo Archives. (2009). WARC Tools Phase III Functional Requirements Specification. London.

HTTrack. (2011). HTTrack Website Copier. Recuperado de http://www.httrack.com/

Hwang, K., Kim, J. K., & Singh, H. (2007). Index to the History of the Web. Cornell University, Computer Science Department.

IIPC. (2011). International Internet Preservation Consorcium. Recuperado de http://netpreserve.org/

IIPC. (2012). IIPC-WarcTools Project. Recuperado de http://netpreserve.org/ projects/warc-tools-project

Internet Archive. (2008). Nutchwax - Home Page. Recuperado de http://archive-access.sourceforge.net

Internet Archive. (2012). About: About the Internet Archive. Recuperado de http://archive.org/about/

Internet Memory Foundation. (2010). Web Archiving in Europe. Recuperado de http://internetmemory.org/images/uploads/Web_Archiving_Survey. pdf

ISO. (2009). ISO. 28500 Information and documentation-WARC file format. Nueva Zelanda.

Jack, P., & Binns, A. (2012). Web Archive - ARC. Recuperado de https://webarchive.jira.com/wiki/display/Heritrix/ARC+File+Format

Library of Congress of USA. (2002). Building a National Strategy for Digital Preservation. Washington, D.C.: Library of Congress of USA.

Lyman, P. (2002). Archiving the World Wide Web. Recuperado de http://www.clir.org/pubs/reports/pub106/web.html

Masanés, J. (2006). Web Archive. New York, USA: Springer-Verlag.

Mohr, G., Stack, M., Ranitovic, I., y Kimpton, D. A. (2004). An Introduction to Heritrix - An open source archival quality web crawler. *4th International Web Archiving Workshop (2004)* (p. 15). Bath, UK: Internet Archive Web Team.

Sanoja, A., León, C., y Torres, G. (2010). Lineamientos para la Construcción de un Archivo Histórico de la Información Digital producida en Venezuela. *CLCAR 2010.Conferencia Latino Americana de Computación de Alto Rendimiento*. Gramado, RS, Brazil: http://gppd.inf.ufrgs.br/clcar2010/program.html.

Smiley, D., & Pugh, E. (2011). Apache Solr 3 Enterprise Search Server. (2a ed.). Birmingham, UK: Packt Publishing Ltd.

Strodl, S., Becker, C., Neumayer, R., y Rauber, A. (2007). How to Choose a Digital Preservation Strategy. *JCDL*. Vancouver, British Columbia, Canada.: ACM.

UNESCO. (2003). Directrices para la preservación del patrimonio digital. Australia: Biblioteca Nacional de Australia, División de la Sociedad de Información de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura.