VNIVERSIDAD NACIONAL
AVTNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE INGENIERÍA
INGENIERÍA EN COMPUTACIÓN

SOFTWARE PARA CONSTRUIR REPOSITORIOS DIGITALES

TESIS PROFESIONAL PARA OBTENER EL TÍTULO DE INGENIERO EN COMPUTACIÓN

ÁREA
BASES DE DATOS

PRESENTA:
VALENCIA VELÁZQUEZ DIEGO IVÁN

DIRECTOR DE TESIS:
MCC. DANTE ORTÍZ ANCONA

CIUDAD UNIVERSITARIA, MÉXICO, ENERO 2013

Software para construir repositorios digitales

Dedicado:

Al Sr. Jesús Valencia Romero
y a la
Sra. Blanca Angélica Velázquez Ramírez

Gracias, por brindarme la oportunidad de conocer desde el inicio hasta el final lo que fue una maravillosa vida escolar, por permitirme conocer a las personas que ahora son mis amigos, por darme su total protección ante todo, por todo el apoyo en las miles de batallas que han librado por mí, por estar conmigo en las derrotas y logros que he obtenido compartiendo conmigo además tristezas y alegrías, por haber respetado mis decisiones por muy erróneas que éstas hayan sido, por mostrarme el camino correcto a seguir de entre todos los que se me han presentado en la vida, por esos sabios consejos que sin ellos estaría perdido y por ayudarme y estar conmigo al dar éste último paso para convertirme en un profesional, pero sobre todo, les agradezco infinitamente por éste inmenso amor en el que me han acobijado y me han permitido crecer, por todo esto y más es que me atrevo a decir que éste es un triunfo más suyo que mío y sólo espero que tengan presente que mis ideales, esfuerzos y logros han sido también suyos e inspirados siempre en ustedes.

Con todo mi respeto, admiración y mi sincero amor…

Padres, Gracias por todo.

AGRADECIMIENTOS

Agradezco sobre todo a mis padres y hermanos por darme su apoyo incondicional en el desarrollo de éste trabajo, así mismo agradezco a mis amigos que siempre han estado presentes para apoyarme en todo y quienes también han servido como fuente de inspiración para el desarrollo de mi trabajo por su invaluable amistad.

También le agradezco a mis honorables sinodales por llevar éste trabajo, con sus excelentes y minuciosas observaciones, a un nivel más profesional y completo en mi perfil como profesional. Por último y con una mención muy especial te agradezco MCC. Dante Ortíz Ancona por haber sido mi tutor y un excelente asesor ayudándome incondicionalmente a la culminación de ésta tesis pues has dado conmigo de los últimos pasos en mi vida como estudiante y has sido para mí un gran maestro y ejemplo a seguir como persona ya que además me has permitido, al conocerte, llevarme una grata experiencia tanto académica, por compartir conmigo un poco de tus extensos conocimientos, como personal pues además de haber sido todo lo ya mencionado, para mí has sido un muy buen amigo.

Gracias por toda tu ayuda Dante.

CONTENIDO

INTRODUCCIÓN

ANTECEDENTES

El uso de bibliotecas digitales es cada vez más notable en numerosos países. Su uso facilita muchos procesos relacionados a la preservación, organización y difusión de material didáctico.

Una biblioteca digital puede ayudarnos a romper muchas barreras (geográficas, culturales, etc.) que impiden la conexión para la colaboración entre personas y así tratar distintos problemas serios de ámbito educativo, como lo es por ejemplo el analfabetismo en comunidades rurales, pues nos ayudan a compartirles información avanzada y actualizada por medio de dispositivos tecnológicos tan útiles como lo son las computadoras, smartphones y toda clase de nueva tecnología moderna que actualmente está desarrollándose, también problemas como las largas distancias que impiden la comunicación con éstas comunidades impidiendo a algunas personas asistir de forma presencial a tomar una clase pueden minimizarse con la implantación de bibliotecas digitales.

También en las comunidades urbanas existen problemas relacionados a los temas de los repositorios digitales y que pueden ser resueltos con el desarrollo de una colección digital. La falta de organización sobre los datos que los usuarios de grandes compañías puedan llegan a tener, la falta de espacio para almacenar documentos físicos y el maltrato o desgaste que los mismos puedan llegar a tener con el paso del tiempo son también problemas que para las sociedades modernas son muy preocupantes.

El compartir información entre distintas organizaciones estudiantiles puede ser una potente herramienta para el crecimiento de un país pues en mi opinión, un país que brinda una buena educación a su gente es un país que sobresale y camina bien y el compartir conocimientos nos brinda una fuente de aprendizaje ilimitada.

Todas estas cuestiones junto con las preocupaciones que numerosas instituciones tanto públicas como privadas de diversos campos (educación, tecnología, empresas, gobierno, cultura, salud, etc.) se plantean sobre sus intereses monetarios y demás, han sido muy discutidas a la hora de pensar en la planeación de la implantación de un repositorio digital que solucione todas las demandas sobre la creación, organización, mantenimiento, manejo, acceso, compartimiento y preservación de colecciones de documentos digitales y que con la creciente evolución tecnológica ya se han estado implantando en gran parte de estas instituciones.

OBJETIVO DE LA TESIS

El objetivo de esta tesis consiste en describir las bases y los principios tecnológicos para la construcción de un repositorio digital así como las características para evaluación de herramientas (de tipo comercial y de licencia libre) dedicadas a la creación de repositorios digitales comparando las diferentes ventajas y desventajas que éstas ofrecen.

ALCANCES DE LA TESIS

Describir los fundamentos tecnológicos que se toman en cuenta para el diseño y desarrollo de un repositorio digital.

Evaluar cuatro herramientas (dos de tipo comercial y dos de licencia libre) dedicadas a la creación de repositorios digitales comparando las diferentes ventajas y desventajas que éstas ofrecen.

ORGANIZACIÓN DE LA TESIS

En el capítulo uno se describen algunos conceptos básicos relacionados a los Archivos, Bases de Datos e Índices, se abordan temas como lo que son los sistemas de archivos así como también sistemas de administración de los mismos. También se habla sobre sistemas de administración de bases de datos (DBMS), índices, tipos de índices y sobre la indización (concepto fundamental para una herramienta de construcción de repositorios digitales).

El capítulo dos trata de explicar el concepto de los metadatos por su importancia en la organización, búsqueda y recuperación de información, así como algunas de sus propiedades, estándares e iniciativas más conocidas sobre los mismos para su uso en bibliotecas digitales, se hace especial alusión en los metadatos de tipo Dublín Core (metadatos utilizados por la mayoría de las herramientas de construcción de repositorios digitales).

El capítulo tres abarca el tema de protocolos de intercambio de información más utilizados puesto que constituyen el mecanismo para compartir información y construir redes de repositorios digitales.

Dentro del cuarto capítulo se abordan temas relacionados a la preservación digital partiendo desde la definición del mismo concepto, algunas problemáticas, se hace

mucho hincapié en algunas estrategias de preservación y también se habla sobre políticas y procedimientos relacionados al tema de la preservación digital.

En el capítulo quinto se hace una evaluación, mediante tablas comparativas, de cuatro de las herramientas más utilizadas para la construcción de repositorios digitales: Greenstone, DSPace, Alfresco y Microsoft SharePoint WorksPace 2012. Cabe mencionar que las dos primeras herramientas son de libre licencia, mientras que las últimas dos son de tipo comercial. Se abordan cuatro conceptos fundamentales de evaluación: evaluación de la instalación, evaluación de la documentación, evaluación de las características técnicas y por último está también la evaluación de las funcionalidades.

En el sexto capítulo se hace una reflexión de los capítulos anteriores dando una perspectiva más general haciendo énfasis sobre la cuestión de qué herramienta es más efectiva de todas las evaluadas.

Para finalizar, se presenta una sección de anexos en la que se muestran ejemplos de repositorios totalmente funcionales (uno por cada herramienta) y se describen para cada uno de ellos algunos de los procesos descritos en los capítulos anteriores referentes al uso de éstas mismas herramientas.

1 ARCHIVOS, BASES DE DATOS E ÍNDICES

1.1 Archivos

Los datos son la materia prima de la que se deriva la información. La información está compuesta por datos que se han recopilado y procesado de manera significativa. Un conjunto de bits se combinan para formar un caracter; los caracteres se unen para representar los valores de los elementos dato, también llamados campos; los elementos dato relacionados entre sí se agrupan para formar registros; los registros que contienen los mismos elementos dato se combinan para formar un archivo (*Long, L. 1995*).

Los archivos se utilizan como un mecanismo para asegurar la permanencia de los datos y cuando el volumen de los datos es bastante significativo para almacenarse en la memoria principal de la computadora, teniendo que almacenarse estos en el disco duro o en algún otro medio de almacenamiento secundario.

Los archivos se clasifican, de acuerdo al contenido en: archivos texto plano, archivos binarios, archivos de datos, archivos de control, archivos de transacción, archivos de registro, archivos de imagen, archivos de sonido, archivos de video, programas, hojas de cálculo, etc.

Un archivo normalmente existe en un directorio y posee un nombre y una extensión, con las cuales es posible identificarlo. Los archivos también poseen un conjunto de atributos que varían de acuerdo al sistema operativo:

- Protección: quién debe tener acceso y de qué forma.
- Contraseña para acceder al archivo.
- Contraseña para editar al archivo.
- Creador del archivo.
- Propietario del archivo.
- Bandera para - lectura / escritura.
- Bandera de visibilidad: 0 visible, 1 para ocultar.
- Bandera de sistema: 0 archivo normal, 1 archivo de sistema.
- Tamaño en bytes de un registro.
- Fecha y hora de creación del archivo.
- Fecha y hora del último acceso al archivo.
- Fecha y hora de la última modificación al archivo.
- Tamaño en bytes en el archivo.
- Tamaño máximo al que puede crecer el archivo.

1.2 Sistemas de Archivos

El sistema de archivos es la estructura y organización de datos en un dispositivo de almacenamiento. En otras palabras, es la manera en que la información es guardada en archivos, unidades de disco y unidades extraíbles (*Hudson, A. 2008*). El sistema de archivos debe contar con alguna interfaz que permite al usuario manipular sus archivos independientemente de la forma como estos se almacenen en la memoria secundaria (*Euán, Avila. 1989*).

A lo largo del tiempo han existido distintos tipos de sistemas de archivos, cada uno de ellos desarrollados para los diversos sistemas operativos que buscaban la manera de mantener su información en un estado persistente. Las especificaciones que podemos encontrar es sumamente variada, algunos por ejemplo permiten la lectura de discos removibles tales como CD's o DVD's, definen el tamaño para una partición y para un archivo, otros además del almacenamiento contienen permisos de acceso así como metadatos correspondientes a los archivos.

1.2.1 Sistemas de Administración de Archivos

Un Sistema de Administración de Archivos (FMS por sus siglas en inglés File Management System) es un sistema formado por un conjunto de archivos y un conjunto de programas para administrarlos. Utiliza los servicios proporcionados por el sistema de archivos del sistema operativo y provee una organización lógica de alto nivel para los datos.

La gestión de los archivos implica tanto la definición de estructuras para el almacenamiento de información como la provisión de mecanismos para la manipulación de la misma.

Desventajas de los FMS:

- Los cambios en los requerimientos requieren cambios en los programas de aplicación.
- Los cambios en las estructuras de almacenamiento requieren cambios en el código de los programas.
- Aumentan los costos de desarrollo.
- Se requiere mayor esfuerzo de programación.
- Frecuentemente se requiere conocer la estructura física de los datos para poder accederlos.

- Comúnmente las aplicaciones no son portables, quedan atrapadas en el equipo en que fueron desarrolladas.
- El mantenimiento de los programas es costoso y complejo.
- Resulta complicado o imposible compartir la información.
- Incorpora mucha redundancia de los datos.
- Fácilmente los datos pueden quedar en un estado inconsistente.
- No provee mecanismos para la integridad de los datos.
- No proporciona mecanismos de seguridad de los datos
- Se tiene que programar el control de la concurrencia de los datos.
- No protege los datos contra fallas del sistema.
- No provee de diccionario de datos.
- No provee de una interfaz de alto nivel con los programadores.

La Tabla 1 muestra algunos ejemplos de Sistemas de Administración de Archivos y como se clasifican en función de cómo procesan la información. Los indizadores de texto construyen índices basándose en el contenido textual de los archivos. Pueden indizar el contenido por palabra o frase y proveen una interfaz de búsqueda para recuperar la información. Podemos realizar una búsqueda por frase o palabra y se recuperan los archivos que contienen dicha frase o palabra (*Ortiz, Dante. 2007*). Los administradores de información gestionan tablas de una base de datos y los índices de dichas tablas.

En la actualidad muchos sistemas de software comercial, libre o propietario siguen utilizando FMS ya sea los mostrados en la tabla 1 o tienen incorporados su propios FMS.

Clasificación	Software
Indizadores de texto	Lucene Zebra Managing Gigabytes Zilverline Lius Regain
Administradores de información	Derby GDBM Berkeley DB

Tabla 1. Ejemplos de Sistemas de Administración de Archivos

1.3 Bases de Datos

Una Base de Datos es una colección de datos interrelacionados y almacenados permanentemente en una computadora tal que:

a. Los datos son compartidos por diferentes usuarios y programas de aplicación, pero existe un mecanismo común para la inserción, actualización, borrado y consulta de los datos.

b. Tanto los usuarios finales como los programas de aplicación no necesitan conocer los detalles de las estructuras de almacenamiento.

1.4 Sistemas de Administración de Bases de Datos

Un DBMS (Data Base Management System - Sistema Administrador de la Base de Datos) es un sistema formado por una Base de Datos y un conjunto de programas para administrarla y explotarla. (*Ortiz, Dante. 2003*)

Servicios proporcionados por un DBMS:

- Almacenamiento y recuperación eficiente de los datos: utilizar mecanismos sofisticados de almacenamiento e indización para recuperar y almacenar en forma rápida la información.
- Minimización de la redundancia de los datos: significa no almacenar datos repetidos o derivados.
- Aseguramiento de la consistencia de los datos: obtener la misma información por peticiones similares en un momento dado.
- Mantenimiento de la integridad de los datos: reglas dictadas por políticas o normas de la empresa y que los datos deben cumplir.
- Otorgamiento de la seguridad de los datos: protección de los datos contra accesos, modificaciones o pérdidas, ya sea en forma intencional o no intencional.
- Control de la concurrencia de los datos: múltiples usuarios pueden accesar a la misma información al mismo tiempo, sin que con ellos se tengan problemas con los datos.
- Protección de los datos contra fallas del sistema: capacidad de restaurar la integridad y consistencia después de una falla del sistema.
- Administración del diccionario de datos: capacidad que da el manejador de la base de datos de poder tener la descripción de los datos que están almacenados en la base de datos.

- Otorgamiento de una interfaz de alto nivel con los programadores: incorporación de un lenguaje como lo es SQL.
- Independencia de datos con respecto a los programas de aplicación: los cambios en la estructura física de almacenamiento no repercuten en los programas de aplicación.

Los usuarios y los programas de aplicación interactúan con la base de datos a través de transacciones. Estas constituyen el mecanismo esencial para asegurar la consistencia y la integridad de los datos.

1.5 Índices

De acuerdo con la Oficina Nacional de Normalización Cuba, un índice es una lista alfabética y sistemática de cualquier concepto o combinación de conceptos que representa el tema específico de un documento y que señala el lugar en que se encuentra cada uno de estos en un documento o en una colección de documentos.

1.6 Indización

De acuerdo a la norma ISO 5963 la indización es la acción de describir o identificar un documento en relación con su contenido. (*Norma Cubana. 1985*)

La indización no concierne a la descripción de un documento como entidad física (por ejemplo, no indica la forma, editor, fecha, etc.), aunque estos factores pueden estar incluidos en un índice de materias si esta información puede permitir a un usuario determinar, de forma más precisa, si un documento dado es relevante o no.
Durante la indización los conceptos se extraen del documento mediante un proceso de análisis intelectual y después se transforman en términos de indización. Tanto el análisis como la transcripción deben realizarse con ayuda de herramientas de indización, como tesauros y sistemas de clasificación.

La indización consiste esencialmente en tres etapas, que tienden a solaparse en la práctica (AENOR, 1991):

a) examen del documento y determinación de su contenido.
La precisión con que se puede examinar un documento depende en gran manera de su forma física. Se pueden distinguir dos casos diferentes: documentos impresos y documentos no impresos.

Los documentos impresos constituyen el material habitual de las bibliotecas y centros de documentación cuyo fondo consiste principalmente en libros, revistas, informes, actas de congresos, etc.
De forma ideal la compresión completa de estos documentos requiere su lectura detallada. Sin embargo, una lectura completa es a menudo impracticable y no siempre necesaria, pero el indizador debe asegurarse de que no se ha descuidado ninguna información útil. Las partes importantes del texto deben examinarse cuidadosamente y se debe prestar especial atención a lo siguiente:

a) título;
b) resumen, si lo tiene;
c) sumario o tabla de contenido;
d) introducción, párrafos iniciales de los distintos capítulos o apartados y conclusiones;
e) ilustraciones, diagramas, tablas y su leyenda o explicación;
f) palabras o frases que están destacadas mediante una tipografía diferente o subrayada.

b) identificación y selección de los conceptos principales del contenido.
Después de examinar el documento el indizador debe identificar las nociones que son elementos esenciales de la descripción del contenido. Las instituciones que patrocinan la realización del índice deben establecer los factores que se consideran importantes en el campo temático cubierto por el índice.
Algunas cuestiones de los criterios a tomar en cuenta son:

a) ¿Trata el documento de algún objeto sometido a una acción?
b) ¿Contiene algún concepto activo? (por ejemplo, una acción, un procedimiento, etc.)
c) ¿Se ve afectado el objeto por la acción identificada?
d) ¿Trata del agente causante de la acción?
e) ¿Se describen los medios para llevar a cabo la acción? (por ejemplo, instrumentos, técnicas o métodos especiales)
f) ¿Existen factores considerados en un medio o lugar particular?
g) ¿Se identifican variables dependientes o independientes?
h) ¿Se trata el tema desde un punto de vista particular no asociado normalmente a ese campo?
(Por ejemplo, estudio de la religión desde un punto de vista sociológico)

c) selección de los términos de indización.
Cuando los conceptos se traducen en términos de indización, el indizador debe observar las reglas siguientes:

a) Los conceptos ya presentes en el lenguaje de indización deben retenerse como descriptores.
b) Los términos que representan nuevos conceptos deben comprobarse, en cuanto a su exactitud y su aceptación, con ayuda de obras de referencia tales como:
- diccionarios y enciclopedias, de autoridad reconocida en la materia en cuestión;
- tesauros
- clasificaciones temáticas.

1.7 Técnicas para crear índices

Aquí se resumen tres técnicas descritas por *Aggarwal, Charu C. (2001)* para la creación de índices en los sistemas de recuperación de información. Aunque Aggarwal demuestra que la técnica de índice Conceptual Efectivo (ICE) es la mejor; los indizadores de texto como Lucene, Zebra y Managing Gigabytes utilizan la técnica de índices invertidos. Sin embargo, muchas herramientas de software para construcción de repositorios digitales incorporan la técnica ICE mediante el uso de los metadatos. Cabe aclarar que los DBMSs no incorporan todavía estas técnicas.

1.7.1 Índice Invertido

La representación de Índice Invertido (II) es el método dominante para indizar texto, pero no es conveniente en búsqueda de similitud entre documentos. El desempeño de la representación II empeora cuando se incrementa el número de palabras en un documento o en los casos en que una palabra tiene una lista invertida demasiado grande.

1.7.2 Índice Semántico Latente (ISL)

Es una técnica para mejorar la calidad en la búsqueda por similitud transformando documentos del conjunto de palabras original a un espacio de conceptos. La idea principal de este método es proyectar los datos en un espacio pequeño, de los datos originales, eliminando los efectos nocivos de sinonimia y polisemia. Trata de minimizar ambigüedad, redundancia y vocabulario sin comprimir la representación. ISL transforma los datos de una representación indizada dispersa (como en II) con dimensionalidad alta a una representación en un espacio real mucho menos

disperso. Desafortunadamente, ISL transforma los datos en un dominio que no es posible brindar técnicas de indizado efectivas.

1.7.3 Índice Conceptual Efectivo (ICE)

Un documento se representa como conjuntos de atributos que corresponden a conceptos con significado. Cada uno de estos conceptos es definido por una palabra con un peso asociado (frecuencia). La palabra con el peso representa a un conjunto de palabras relacionadas semánticamente. La representación ICE es una representación comprimida que reduce ambigüedad, redundancia y vocabulario no relacionado en un documento. Un vez que se reduce la dimensión de un documento se utiliza el método de índices invertidos para indizar los documentos.

El método ICE es mucho mejor que II en búsquedas de similitud y preserva la misma calidad de los resultados, tiene una gran eficiencia computacional y de almacenamiento. Por ejemplo, en una muestra de 167,000 documentos se requirió 87.7 Mb usando el método II y 8.3 Mb usando ICE para el almacenamiento de los índices.

1.8 Índices en Bases de Datos

Un índice es un archivo auxiliar que proporciona acceso rápido a los registros de un archivo de datos. El índice utiliza un valor destino del campo de ordenamiento para obtener las direcciones de disco de los registros deseados.

Se sabe que existen dos organizaciones que subdividen un archivo grande en un conjunto de minicúmulos (secuenciales indizados y verificados) para reducir el costo de acceso a disco para una búsqueda dada, en comparación con una organización secuencial, una descomposición en N minicúmulos reduce en 1/N éste costo.

Una solución intermedia es que el DBMS debe mantener un arreglo de memoria de N direcciones de bloque, que no puede crecer de manera indefinida por limitaciones de la memoria. En consecuencia, los minicúmulos individuales se convierten en una última instancia en cadenas grandes de bloques en sí mismas, lo cual compromete el rendimiento dentro de cada minicúmulo. (*Johnson, J. 2000*)

Una solución obvia almacena el arreglo de direcciones de bloque en el disco junto con los bloques de datos; entonces el DBMS los numera parcialmente en la

memoria según sea necesario. Un archivo auxiliar de esta naturaleza se llama índice, y sus registros contienen valores clave y direcciones correlacionadas de bloque. En algunos casos, las direcciones de bloque que se refieren al archivo principal de datos indican dónde es posible hallar un registro objetivo. En otros casos, las direcciones de bloque se refieren a otras porciones del índice, y hay necesidad de explorar más para hallar la referencia a los datos. Cuando se determine el costo de lectura de un registro, por supuesto que deben contarse los accesos a disco de más que leen el índice. En consecuencia, es deseable organizar el archivo de índice para reducir al mínimo este gasto.

1.9 Tipos de índices

1.9.1 Índices Dispersos de llaves no Duplicadas

Un índice jerárquico disperso acelera el acceso a un archivo secuencial indizado. Utiliza el mismo campo, o grupo de campos, como la organización secuencial indizada: el campo de ordenamiento del archivo. Un índice jerárquico disperso recibe ese nombre porque exhibe una estructura jerárquica y puede catalogar todo el archivo de datos al rastrear sólo un subconjunto de los valores del campo de ordenamiento, además de que no contiene una entrada por cada registro del archivo. Para localizar un registro se busca la entrada del índice con el valor más grande que sea menor o igual al valor buscado. Un índice jerárquico disperso a veces se llama índice ISAM, por sus siglas en inglés, que significan método de acceso secuencial indizado.

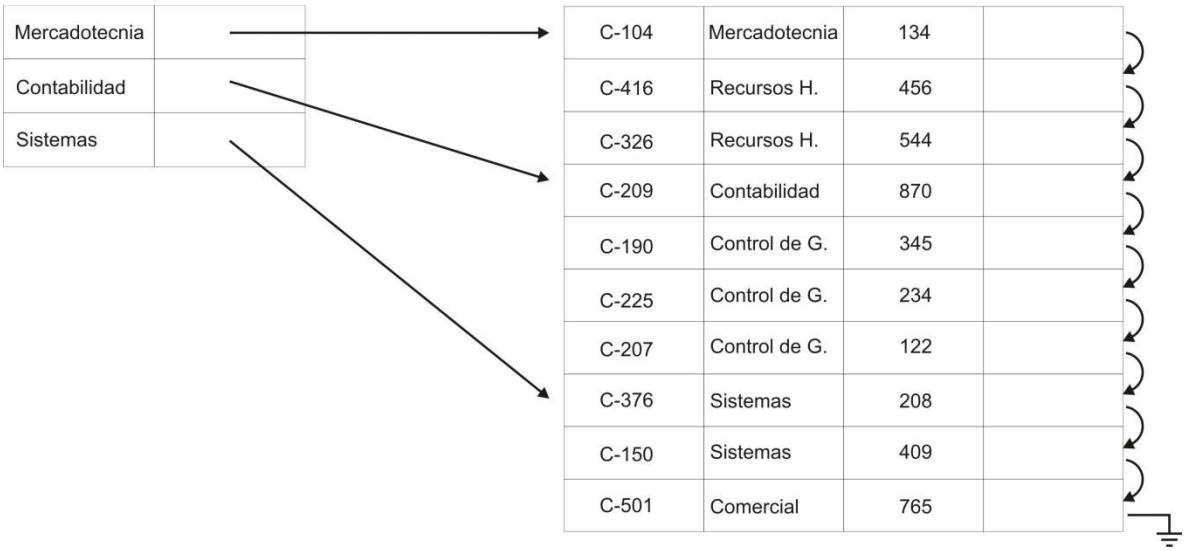


Figura 1. Ejemplo de índices dispersos

1.9.2 Índices Densos

En estos tipos de índice aparece un registro índice por cada valor de la clave de búsqueda en el archivo. El registro índice contiene el valor de la clave y un apuntador al primer registro con ese valor de la clave de búsqueda. El resto de registros con el mismo valor de dicha clave se almacenan consecutivamente después del primer registro, dado que, el índice es primario, los registros se ordenan sobre la misma clave de búsqueda. (Silberschatz, 2002)

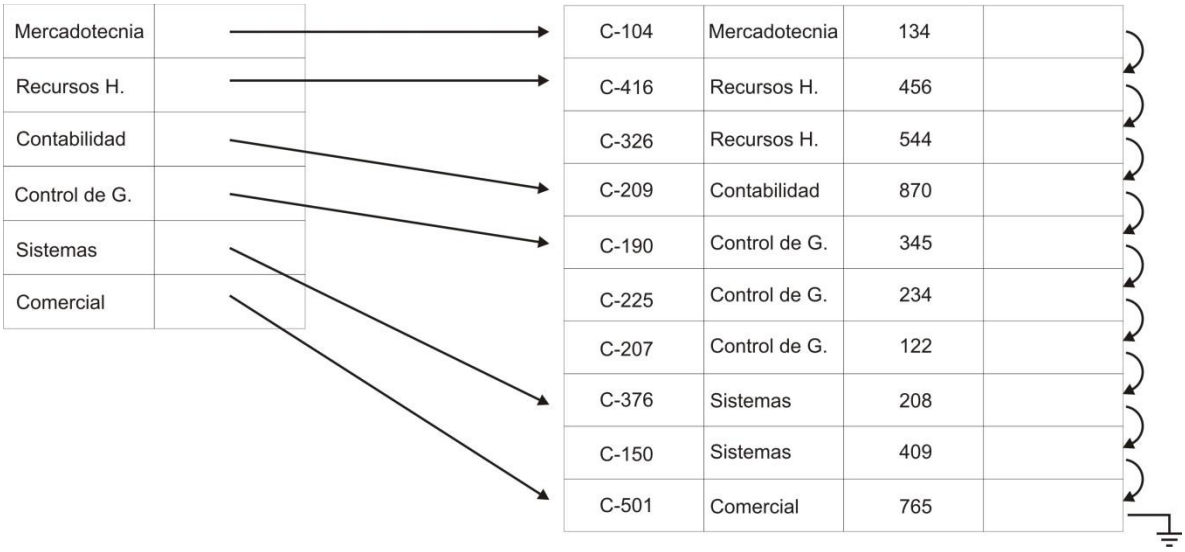


Figura 2. Ejemplo de índices densos

1.9.3 Árboles B

Un índice de árbol B es una generalización del índice jerárquico denso, pero con la capacidad para reorganizarse dinámicamente a sí mismo para evitar estructuras desequilibradas. Con sus diversas variantes, la familia de estructuras de árbol B es en la actualidad el método más popular para poner en práctica índices en sistemas de administración de bases de datos.

1.9.4 Propiedades de definición de un árbol B+

Los bloques de índice de un árbol B+ son los nodos, y el bloque individual de nivel superior es la raíz. Los nodos de un nivel dado son los nodos hijo de un padre en el nivel inmediatamente precedente. El nivel inferior está formado por nodos hoja, que contienen las direcciones de registro de la información. Un índice de árbol B+ es entonces un árbol con las siguientes restricciones:

1. Cada nodo contiene un conjunto ordenado de llaves y apuntadores de la forma ($p_0^{(q)}$, $k_1^{(q)}$, $p_1^{(q)}$, $k_2^{(q)}$, $p_2^{(q)}$, …, $k_t^{(q)}$, $p_t^{(q)}$), donde $k_1^{(q)} < k_2^{(q)} < \ldots < k_t^{(q)}$. El exponente da la profundidad del nodo en el árbol, teniendo la raíz una profundidad de uno. Por cada nodo que no sea hoja, las llaves definen una secuencia de espacios de búsqueda que no se traslapa.
2. El número de llaves de un nodo no puede exceder un valor máximo, n, el orden del árbol B+. El orden más pequeño para un árbol B+ es dos, pero por lo general n es mucho mayor.
3. Sea t el número de llaves en un nodo arbitrario que no sea la raíz. Así [n/2] ≤ t ≤ n. Para la raíz, 1 ≤ t ≤ n, a menos que el árbol se encuentre vacío.
4. Todas las rutas desde la raíz hasta nodo hoja son de longitudes iguales.

Una estructura de árbol B+ mantiene llenos todos sus nodos, excepto la raíz, por lo menos hasta la mitad de su capacidad. También mantiene todos los nodos hoja a una distancia uniforme de la raíz. Los espacios de búsqueda en una ruta de raíz a hoja forman una secuencia anidada de intervalos de tamaño decreciente.

Sean Pi y Ki apuntadores y valores clave de búsqueda ordenados respectivamente.

Para Pi:
- Si el nodo es interno o raíz: se apunta a otro nodo de un nivel inferior
- Si el nodo es hoja: se apunta a un registro o a un cajón de apuntadores a registros

Y sea n = 3 que de fine el valor máximo de apuntadores a n - 1 valores clave de búsqueda.

Nodos internos: tienen entre un mínimo de (n/2) o bien y un máximo de n hijos (apuntadores).

Nodos hojas: tienen almenos (n-1) / 2 valores de claves (Ki) y máximo (n - 1) valores de Ki.

Nodo raíz: tiene entre 1 y n hijos o apuntadores.

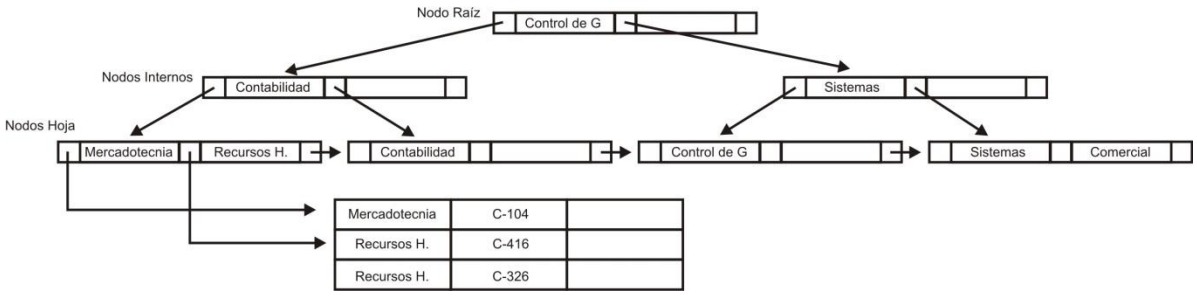


Figura 3. Ejemplo de árboles B+

1.10 Software libre para indización, búsqueda y recuperación de información

Según *Ortíz, Dante (2007)* las herramientas de software libre para indización, búsqueda y recuperación de información más comunes son Lucene [1], Zebra [2] y Managing Gigabytes [3]. Todas estas herramientas tienen como núcleo la representación de índices invertidos (*véase figura 4*). Managing Gigabytes incorpora un núcleo de nivel más bajo que maneja algoritmos bastante sofisticados para comprimir la información y manejar de manera más eficiente imágenes, audio y video.

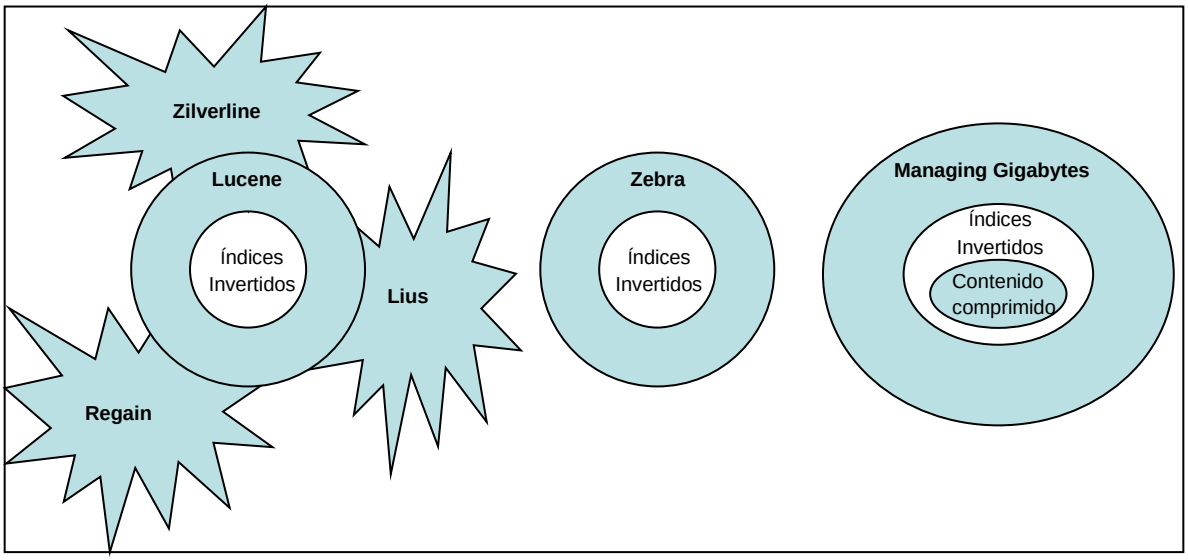


Figura 4. Software libre para indizado, búsqueda y recuperación de información

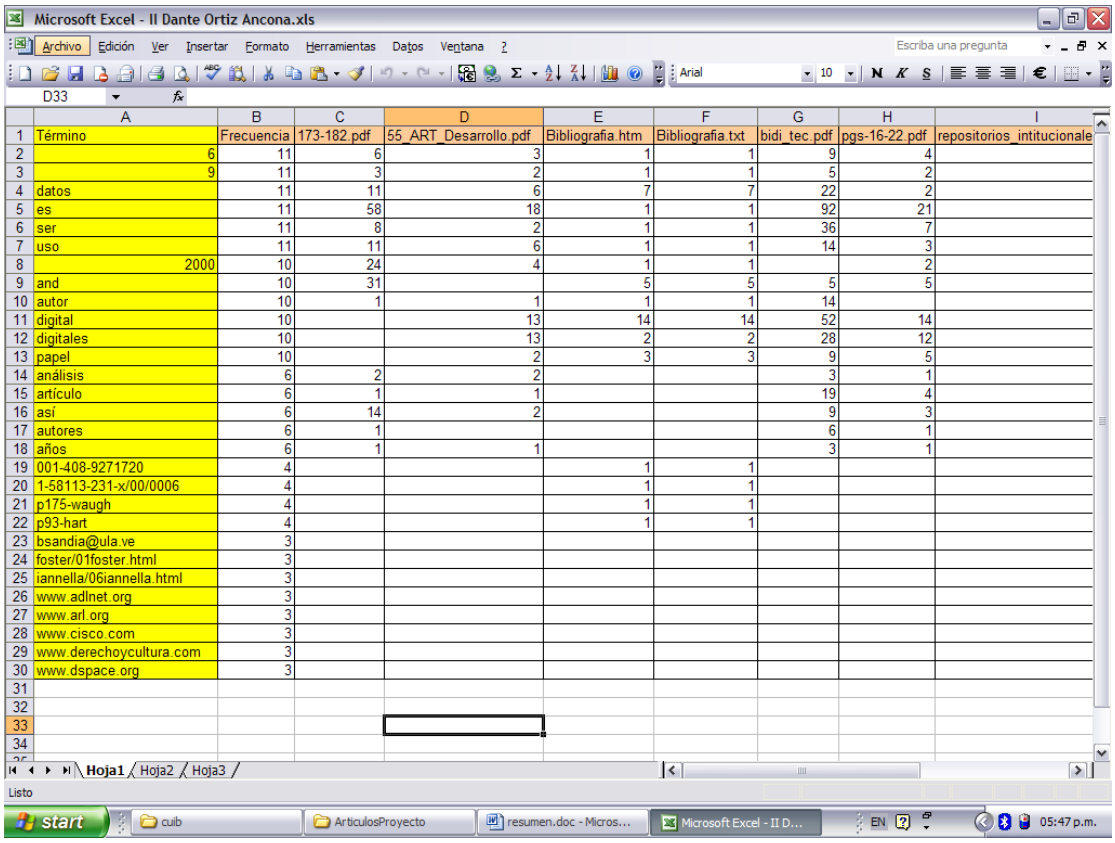
Lucene es una interfaz para programas de aplicaciones que contiene un motor para indizar, buscar y recuperar información tanto de registros como texto completo. Es sin duda, dentro de su clase, el software con mayor respaldo en soporte, documentación y desarrollo de proyectos. Aunque utiliza principalmente el idioma inglés, provee una interfaz de programación que le permite incorporar, con gran facilidad, otros idiomas. Fue desarrollado en el lenguaje de programación Java, esto le ha permitido una gran portabilidad para funcionar en cualquier sistema operativo y arquitectura de cómputo, así como una gran interoperabilidad con otros sistemas computacionales. No utiliza un sistema de metadatos descriptivo, sin embargo, resulta bastante simple adaptarle cualquier sistema de metadatos.

Lucene utiliza por defecto un analizador lexicográfico, para texto en idioma inglés, eliminando del vocabulario palabras sin importancia en búsqueda de información (stopwords) tales como artículos, preposiciones, verbos comunes (is, have, get, etc.), palabras acerca de la estructura del texto, etcétera. El analizador sintáctico permite que un usuario o programador especifique un archivo que contenga esta lista de palabras dando la posibilidad de aumentar o disminuir el diccionario de palabras sin importancia. Además reconoce términos con las características siguientes: secuencias de letras y dígitos (p175waugh), apóstrofos (O'Reilly), acrónimos (H.P.), compañías (AT&T), direcciones de correo electrónico (dante@dgb.unam.mx), nombres de servidores WEB (www.dspace.org), números seriales (direcciones IP como 132.248.9.31, números de punto flotante como 3.1416). Puede desarrollarse un analizador lexicográfico para el idioma español o cualquier otro idioma (que quizá incorpore el uso de raíces de palabras y/o tesauros) y utilizarlo en lugar del que se tiene por defecto. Por ejemplo, sea "vacias.txt" el archivo que contiene las palabras irrelevantes para la búsqueda y sea "SpanishAnalizer" el Analizador lexicográfico para el idioma español, entonces, el fragmento de código, en el lenguaje de programación Java, para construir dicho analizador lexicográfico sería: new StandardAnalyzer(new File("vacias.txt").

La figura 5 muestra una representación en Excel de un subconjunto, del índice construido, utilizando el analizador lexicográfico en idioma español anteriormente mencionado. La primera columna contiene la lista de términos, el primer renglón ilustra parte de una lista con un total de 11 identificadores de documentos de texto en diversos formatos. La segunda columna ilustra la cantidad de documentos en que aparece el término. La celda de intersección entre el término y el identificador del documento muestra el número de veces que aparece el término en el documento de texto. Por simplicidad no se muestran las posiciones del término dentro del documento de texto.

Aunque el motor de Lucene fue desarrollado para indizar, buscar y recuperar información en texto plano. Se han desarrollado otras herramientas de software libre tales como Zilverline [4], LIUS (Lucene Index Update and Search) [5] y Regain [6] que tienen como núcleo a Lucene (*véase figura 4*) y que amplían su funcionalidad al incorporar filtros que permiten transformar documentos de diferentes formatos (Word, Powerpoint, Excel, Postscript, PDF, HTML, XML, etc.) a texto plano.

Estas herramientas proveen una interfaz de usuario vía WEB para administrar el índice, incorporar documentos de texto, realizar búsquedas avanzadas (incorporando operadores lógicos, de agrupamiento, de selección de campos, comodines, de proximidad y de rangos) y recuperar información ya sea textual o descriptiva.



Microsoft Excel - II Dante Ortiz Ancona.xls

	Término	Frecuencia	173-182.pdf	55_ART_Desarrollo.pdf	Bibliografia.htm	Bibliografia.txt	bidi_tec.pdf	pgs-16-22.pdf	repositorios_intitucionale
2	6	11	6	3	1	1	9	4	
3	9	11	3	2	1	1	5	2	
4	datos	11	11	6	7	7	22	2	
5	es	11	58	18	1	1	92	21	
6	ser	11	8	2	1	1	36	7	
7	uso	11	11	6	1	1	14	3	
8	2000	10	24	4	1	1		2	
9	and	10	31		5	5	5	5	
10	autor	10	1	1	1	1	14		
11	digital	10		13	14	14	52	14	
12	digitales	10		13	2	2	28	12	
13	papel	10		2	3	3	9	5	
14	análisis	6	2	2			3	1	
15	artículo	6	1	1			19	4	
16	así	6	14	2			9	3	
17	autores	6	1				6	1	
18	años	6	1	1			3	1	
19	001-408-9271720	4			1	1			
20	1-58113-231-x/00/0006	4			1	1			
21	p175-waugh	4			1	1			
22	p93-hart	4			1	1			
23	bsandia@ula.ve	3							
24	foster/01foster.html	3							
25	iannella/06iannella.html	3							
26	www.adlnet.org	3							
27	www.arl.org	3							
28	www.cisco.com	3							
29	www.derechoycultura.com	3							
30	www.dspace.org	3							

Figura 5. (Representación en Excel de un subconjunto del índice invertido, generado con un analizador lexicográfico del idioma español)

Zebra es una interfaz para programas de aplicaciones que contiene un motor para indizar, buscar y recuperar información. Es una herramienta de propósito general y de rendimiento alto, indiza texto estructurado, lee registros en una variedad de formatos de entrada (correo electrónico, XML, MARC) proporcionando acceso a ellos a través de una poderosa combinación de expresiones de búsquedas lógicas y de relevancia. Soporta bases de datos grandes (decenas de millones de registros, decenas de gigabytes de datos) permitiendo actualizaciones seguras en tiempo real. Soporta el protocolo estandarizado Z39.30 para recuperación e intercambio de información. Cuenta con un amplio respaldo en soporte, documentación (Hammer S. 2005) y desarrollo. Fue desarrollado en el lenguaje de programación C estándar, esto le ha permitido una gran portabilidad para funcionar en cualquier sistema operativo y arquitectura de cómputo, así como

interoperabilidad con otros sistemas computacionales. Provee una interfaz en modo de comandos para administrar el índice y para búsqueda y recuperación de información, permitiendo realizar búsquedas avanzadas (incorporando operadores lógicos, de agrupamiento, de selección de campos, comodines y proximidad).

Managing Gigabytes es una interfaz para programas de aplicaciones que contiene un motor para indizar, buscar y recuperar información de texto completo, archivos binarios, imágenes pictóricas o textuales. Tiene muy poco respaldo en soporte y la única documentación es el libro de *Witten Ian H. (1999)*. A diferencia de otras herramientas, provee algoritmos bastante sofisticados para comprimir texto e imágenes. Fue desarrollado en el lenguaje de programación C y puede interactuar con otros sistemas computacionales. No utiliza un sistema de metadatos descriptivo. Provee un diccionario, en idioma inglés, de palabras sin importancia en búsqueda de información (stopwords) y manejo de raíces de palabras (stemming) para aumentar relevancia en recuperación de documentos y reducir la dimensión del índice. Tiene implantados métodos estadísticos para clasificación y organización de información. Permite realizar búsquedas avanzadas (incorporando operadores lógicos, de agrupamiento, de selección de campos, comodines y proximidad) y recuperación de información de forma interactiva y distribuida.

2 METADATOS

2.1 Definición de Metadatos

El término metadatos describe varios atributos de los objetos de información y les otorga significado, contexto y organización, es decir, un metadato no es más que un dato estructurado sobre la información, o bien, información sobre información, de forma más simple se dice que los metadatos son datos sobre datos.

Los metadatos en el contexto de la Web, son datos que se pueden guardar, intercambiar y procesar por medio del ordenador y que están estructurados de tal forma que permiten ayudar a la identificación, descripción clasificación y localización del contenido de un documento o recurso web y que, por tanto, también sirven para su recuperación. (*Lamarca, María. 2011*)

2.2 Origen de los Metadatos

Debido a la gran diversidad y volumen de las fuentes y recursos en Internet, se hizo necesario establecer un mecanismo para etiquetar, catalogar, describir y clasificar los recursos presentes en la World Wide Web con el fin de facilitar la posterior búsqueda y recuperación de la información. Este mecanismo los constituyen los llamados metadatos.

El concepto de metadatados -datos sobre datos- se puede entender en un sentido amplio o en un sentido más estricto. Por ejemplo, en un sentido amplio, si entendemos que metadatos es un término que se utiliza para describir datos que ofrecen el tipo y la clase de la información, esto es, son datos acerca de datos, podemos considerar que el catálogo de una biblioteca o un repertorio bibliográfico son tipos de metadatos. Estos tipos de metadatos emplean, fundamentalmente, reglas de catalogación y formatos para transmitir la información, como los formatos MARC. Así considerados, cada ficha catalográfica es un conjunto de metadatos de un libro o bien de un autor y los metadatos proporcionan una información básica sobre las obras de un autor y lo relacionan con otras obras del mismo autor u otras obras de similar contenido. De la misma forma, los registros de una base de datos llevada a cabo para indizar o hacer un resumen documental, podrían también considerarse como metadatos.

Sin embargo, si acotamos la definición de metadatos dándole un sentido más estricto, los metadatos sólo serían posibles en un contexto digital y en red ya que sólo dentro de este contexto se pueden utilizar los metadatos con la función que

les caracteriza, que es la de la localización, identificación y descripción de recursos, legibles e interpretables por máquina.

2.3 Clasificación de Metadatos

Existen distintos modelos de metadatos, cada uno de ellos con distintos esquemas de descripción. En los distintos modelos, cada objeto se describe por medio de una serie de atributos y el valor de estos atributos es el que puede servir para recuperar la información. Dependiendo de la clase de metadatos puede existir: información sobre elementos de datos o atributos, información sobre la estructura de los datos, información sobre un aspecto concreto, etc. De forma general, podemos encontrar metadatos referidos a:

- el contenido (concepto)
- aspectos formales (tipo, tamaño, fecha, lengua, etc.)
- información del copyright
- información de la autentificación del documento o recurso
- información sobre el contexto (calidad, condiciones o características de acceso, uso, etc.)

Con fines prácticos, los tipos y las funciones de los metadatos pueden clasificarse en tres amplias categorías: descriptivos, estructurales y administrativos. Estas categorías no siempre tienen límites bien definidos y con frecuencia presentan un significativo nivel de superposición. Por ejemplo, los metadatos administrativos pueden incluir una amplia gama de información que podría ser considerada como metadatos descriptivos y estructurales (*Library, cornell. 2003*)

Cabe mencionar también que existen las siguientes iniciativas de metadatos:

- Metadatos para la descripción
- Metadatos para presentaciones
- Metadatos para la industria y el comercio electrónico
- Metadatos para multimedia
- Metadatos para la educación y el aprendizaje
- Metadatos para el gobierno y la administración
- Metadatos geoespaciales
- Metadatos generales

2.4 Estándares de Metadatos para Repositorios Digitales

2.4.1 Metadatos Descriptivos

Su objetivo es la descripción y la identificación de recursos de información. A nivel sistema local nos permiten la búsqueda y la recuperación (por ejemplo, búsqueda de una colección de imágenes para encontrar pinturas con ilustraciones de animales).

Algunos ejemplos de estos tipos de metadatos son:

- Identificadores únicos (Handle, DOI)
- Atributos Físicos (medios, condición de las dimensiones)
- Atributos Bibliográficos (título, autor/creador, idioma, palabras claves)

2.4.2 Metadatos Administrativos

Tienen como objetivo facilitar la gestión y procesamiento de las colecciones digitales tanto a corto como a largo plazo.

Incluyen también datos técnicos sobre la creación y el control de calidad, la gestión de derechos y requisitos de control de acceso y utilización, además de información sobre acción de preservación.

Algunos ejemplos de estos tipos de metadatos son:

- Datos técnicos tales como tipo y modelo de escáner, resolución, profundidad de bit, espacio de color, formato de archivo, compresión, fuente de luz, propietario, fecha del registro de derecho de autor, limitaciones en cuanto al copiado y distribución, información sobre la licencia, actividades de preservación (ciclos de actualización, migración, etc.).

2.4.3 Metadatos de Preservación

El Diccionario de Datos PREMIS define los metadatos de preservación como la información que un repositorio utiliza para llevar a cabo el proceso de preservación digital (*véase capítulo 4 para más información*). En concreto, el grupo se centró en los metadatos destinados al mantenimiento de la viabilidad, la disponibilidad, la claridad, la autenticidad y la identidad en el contexto de la preservación. Por lo tanto, los metadatos de preservación engloban una serie de categorías de

metadatos que normalmente se usan para diferenciarlos: administrativos (incluidos derechos y permisos), técnicos y estructurales. Se prestó especial atención a la documentación sobre la procedencia digital (la historia de los objetos) y a la documentación de las relaciones, especialmente aquellas entre distintos objetos dentro del repositorio de preservación. (*BNE. n.d.*)

2.5 Creación de metadatos y Formas de asignarlos a recursos digitales

En cuanto a la creación de los metadatos hay que tener en cuenta dos aspectos importantes: quién los asigna y cómo los asigna.

En la inmensa maraña de la Web, la mayor parte de los documentos y recursos digitales son creados por autores personales individuales sin ninguna experiencia en lenguajes hipertextuales, metadatos, etc. puesto que existen sistemas de gestión de hipertexto y herramientas que permiten la edición de páginas web de forma muy sencilla. Sin embargo, los diseñadores y editores web son cada vez más conscientes de que si quieren tener verdadera presencia en la red, deben facilitar que los buscadores, los robots y agentes inteligentes sean capaces de indizar sus páginas para que los usuarios puedan encontrarlas y recuperarlas. Desde hace algunos años, los principales editores web permiten la inclusión de etiquetas META, ya que los buscadores rastrean en la cabecera de los documentos para extraer las etiquetas <META NAME="KEYWORDS"> y <META NAME="DESCRIPTION"> e indizar las páginas.

Tanto las grandes instituciones y empresas, así como las bibliotecas y repositorios digitales cuentan ya con personal especializado capaz de indizar y catalogar sus propios recursos digitales y entre estos, cada vez se va extendiendo más el uso de metadatos en sus distintas modalidades -desde la simple utilización de etiquetas META hasta la utilización de estándares de metainformación o el empleo de metadatos muy ricos y especializados.

En cuanto a la forma de asignar metadatos, existen varios modos de asociar metadatos con recursos digitales:

- Incrustando los metadatos dentro del propio documento:

Esto implica que los metadatos deben ser creados al mismo tiempo que se crea el recurso, a menudo por el autor. Generalmente se almacenan embebidos y codificados en la cabecera del documento y eso permite que esta metainformación sea transportada por el sistema a la vez que se transporta el contenido del documento.

- Asociando los metadatos:

por medio de archivos acoplados a los recursos a los que describen. La ventaja de los metadatos asociados se deriva de la facilidad relativa de poder manejar los metadatos sin cambiar el contenido del recurso en sí mismo. Estos metadatos persisten aunque el documento ya no esté accesible. Para su indización es preciso contar con herramientas específicas. Este tipo de metadatos se utiliza, sobre todo, para material multimedia, imágenes, etc. Una forma sencilla de crear metadatos asociados es a través del elemento LINK de HTML.

- Metadatos independientes:

Los metadatos se mantienen en un depósito separado, generalmente una base de datos mantenida por una organización que puede o no tener el control directo o tener acceso al contenido del recurso. De esta forma, es mucho más fácil gestionar tanto los metadatos como los recursos -almacenarlos, mantenerlos, actualizarlos, convertirlos a otros formatos, etc.- y, además, es posible que múltiples conjuntos de metadatos pueden referirse al mismo recurso. Este es el método que suelen emplear muchas organizaciones para que sus datos no sean públicos, ya que de esta forma permanecen inaccesibles a los motores de búsqueda.

Además, la creación de metadatos puede realizarse de forma manual o de forma automática, o bien, mediante una combinación de ambos métodos. Los primeros programas que servían para construir hipertexto o elaborar páginas web sólo permitían navegar, ver, distribuir y enlazar las páginas mediante enlaces. Actualmente, la mayor parte de estos programas cuentan también con las herramientas necesarias para generar de forma automática tablas de contenido, para indizar páginas y para añadir metadatos.
Pero al margen de los tradicionales editores web, existen también una serie de herramientas y aplicaciones que permiten crear tanto metaetiquetas, como crear metadatos con distintos niveles de metainformación y funcionalidades. Hay programas muy sofisticados que son capaces de gestionar conocimiento,

herramientas para crear tesauros, aplicaciones para desarrollar ontologías, mapas temáticos, etc; pero también existen herramientas mucho más sencillas cuya función primordial es crear metaetiquetas y metadatos.

2.6 Metadatos Dublín Core

La Iniciativa Dublin Core (DCMI) comenzó en 1995 en un encuentro en Dublin, Ohio (USA) en el que participaron el NCSA (National Center for Supercomputing Applications) y OCLC (On Line Library Computer Center), junto con representantes de la IETF (Internet Engineering Task Force) y en el que bibliotecarios, proveedores de contenido y expertos en lenguajes de marcado pretendieron desarrollar estándares para describir los recursos de información y facilitar su recuperación.

Hoy, los metadatos Dublin Core se han convertido en uno de los estándares más extendidos para la recuperación de información en la World Wide Web y el DC se ha convertido en un vocabulario muy utilizado no sólo en el ámbito bibliotecario y documental, sino en otros muchos sectores.

El conjunto de elementos Dublin Core se centró en 13 elementos, pero concluyó con 15 descriptores como resultado de un consenso y un esfuerzo interdisciplinar e internacional. Ya existen transcripciones a 20 idiomas y ha sido adoptado por el CEN/ISS (European Committee for Standardization / Information Society Standardization System) y es también estándar oficial del WWW Consortium. Los metadatos Dublin Core son utilizados como base tanto por gobiernos como por agencias supranacionales y muchas otras iniciativas de metadatos pertenecientes a comunidades específicas como bibliotecas, archivos, en educación, negocios, etc.

Los metadatos Dublin Core tratan de ubicar, dentro de Internet, los datos necesarios para describir, identificar, procesar, encontrar y recuperar un documento introducido en la red. Si este conjunto de elementos Dublin Core se lograra aceptar internacionalmente supondría que todos los procesos que indizan documentos en Internet encontrarían, en la cabecera de los mismos, todos los datos necesarios para su indización y además estos datos serían uniformes. Si el Dublin Core lograra estandarizar los metadatos de la cabecera de los documentos se facilitaría su indización automática y mejoraría la efectividad de los motores de búsqueda.

2.6.1 Estándares DCMI y Especificaciones Dublín Core

Entre los estándares DCMI y especificaciones DC figuran las siguientes:

- Codificación Dublin Core en HTML (IETF RFC 2731): http://www.ietf.org/rfc/rfc2731.txt
- Metadatos Dublin Core para la Recuperación de Recursos. (IETF RFC 2413). http://www.ietf.org/rfc/rfc2413.txt
- Conjunto de Elementos de metados Dublin Core, Versión 1.1: Descripción de Referencia (Dublin Core Metadata Element Set, Version 1.1: Reference Description): http://Dublin_Core.org/documents/dces/ Describe el conjunto de los 15 elementos Dublin Core.
- Términos de metadatos Dublin Core (DCMI Metadata Terms) http://Dublin Core.org/documents/dcmi-terms/ Recoge todos los términos de metadatos utilizados por la Iniciativa de Metadatos Dublin Core, incluyendo elementos, elementos refinados, esquemas de codificación y términos del vocabulario.
- Vocabulario Tipo DCMI (DCMI Type Vocabulary): http://Dublin Core.org/documents/dcmi-type-vocabulary/ El Vocabulario Tipo DCMI proporciona una lista general de términos aprobados que pueden usarse como valores por el elemento Recurso Tipo para identificar el género del recurso.
- Calificadores Dublin Core (Dublin Core Qualifiers): http://Dublin Core.org/documents/dcmes-qualifiers/ Este documento que, en principio, describía los los calificadores principales que regían Dublin Core, las dos categorías de calificadores y ejemplos de listas de calificadores aprobados por la Comisión de Uso de Dublin Core, ahora remite a DCMI Metadata Terms http://DublinCore.org/documents/dcmi-terms/ pues en esta Especificación es donde quedan recogidos tanto los elementos para refinar, como los esquemas de codificación, los dos tipos de calificadores de los elementos Dublin Core.
- Esquema de codificación del punto DCMI (DCMI Point Encoding Scheme). http://Dublin Core.org/documents/dcmi-point/ Un punto de localización en el espacio, y métodos para codificarlos en una cadena de texto.
- Método de Codificación Periódica de DCMI (DCMI Period Encoding Scheme): http://Dublin_Core.org/documents/dcmi-period/ Para especificar los límites de un intervalo de tiempo y los métodos para codificar éste en una cadena de texto.
- DCMI DCSV (Dublin Core Structured Values) Sintaxis para escribir una lista de valores etiquetados en una cadena de caracteres: http://Dublin Core.org/documents/dcmi-dcsv/ Se describe un método para grabar una

lista de valores etiquetados en una cadena de caracteres, llamado Valores Estructurados Dublin Core, con la etiqueta DCSV. El propósito de esta anotación es ofrecer información estructurada en las descripciones de metadatos Dublin Core.

- Esquema de Codificación DCMI (DCMI Box Encoding Scheme): http://Dublin Core.org/documents/dcmi-box/ Especificación de los límites espaciales de un lugar y métodos para codificarlo en una cadena de texto.
- Usar Dublin Core (Using Dublin Core): http://Dublin Core.org/documents/usageguide/ Este documento es un punto de acceso para los usuarios de Dublin Core, tanto para no especialistas, a quienes les ayudará para la creación de registros descriptivos simples para fuentes de información; como para especialistas, que encontrarán un punto de referencia útil para la documentación de Dublin Core, con sus cambios y ampliaciones.
- Política de Espacios de nombre DCMI (Namespace Policy for the Dublin Core Metadata Initiative): http://Dublin Core.org/documents/dcmi-namespace/ Un namespace XML es una colección de nombres, identificados por una referencia URI, que son usados en documentos XML como elementos tipo y atributos de nombre. El uso de los namespace XML para identificar excepcionalmente términos de metadatos, permite a esos términos no ser utilizados de manera ambigua. DCMI adopta este mecanismo para la identificación de todos los términos DCMI. Este documento especifica los acuerdos realizados para identificar actuales y futuros namespaces DCMI.
- Expresar Dublin Core en meta-elementos HTML/XHTML y elementos de enlace (Expressing Dublin Core in HTML/XHTML meta and link elements) http://Dublin Core.org/documents/dcq-html/ Este documento describe cómo usar metadatos Dublin Core codificados en elementos <meta> y <link>.de HTML/XHTML.
- Expresar Simple Dublin Core en RDF/XML (Expressing Simple Dublin Core in RDF/XML): http://Dublin Core.org/documents/dcq-html/ El formato Dublin Core puede ser representado en muchos formatos de sintaxis. Este documento explica cómo codificar DCMES en RDF/XML, ofrece una DTD para validar los documentos y describe un método para enlazarlos desde las páginas web.
- Guía para implementar Dublin Core en XML (Guidelines for implementing Dublin Core in XML): http://Dublin Core.org/documents/dc-xml-guidelines/ Este documento ofrece una guía para implementar aplicaciones de metadatos Dublin Core usando XML, tanto en aplicaciones DC simples como calificadas.

- Declaraciones de términos Dublin Core representadas en lenguaje de esquemas XML (DCMI term declarations represented in XML schema language): http://Dublin_Core.org/schemas/xmls/ Este documento muestra los esquemas XML que utilizan metadatos Dublin Core.

2.6.2 Clasificación de elementos DC

Podemos clasificar el conjunto de elementos Dublin Core en 3 grupos que indican la clase o el ámbito de la información que contienen:

Elementos relacionados principalmente con el contenido del recurso:

Title (titulo)
Subject (tema)
Description (descripción)
Source (fuente)
Languaje (lenguaje)
Relation (relación)
Coverage (cobertura).

Elementos relacionados principalmente con el recurso cuando es visto como una propiedad intelectual:

Creator (autor)
Publisher (editor) y, otras colaboraciones
Contributor (otros autores/colaboradores)
Rights (derechos).

Elementos relacionados principalmente con la instanciación del recurso:

Date (fecha)
Type (tipo de recurso)
Format (formato)
Identifier (identificador

En la tabla 2 se muestra una breve descripción de cada uno de los metadatos de Dublin Core y la etiqueta correspondiente según la norma estandarizada.

Elemento Dublín Core	ETIQUETA DEL ELEMENTO DC. http://Dublin Core.org/documents/dcmi-terms/	DESCRIPCIÓN
Title (titulo)	**DC. Title** http://purl.org/dc/elements/1.1/title	**Título: El nombre dado a un recurso.** Típicamente, un título es el nombre formal por el que es conocido el recurso.
Creator (autor)	**DC. Creator** http://purl.org/dc/elements/1.1/creator	**Autor: La entidad primariamente responsable de la creación del contenido intelectual del recurso.** Entre los ejemplos de un creador se incluyen una persona, una organización o un servicio. Típicamente, el nombre del creador podría usarse para indicar la entidad.
Subject (tema)	**DC. Subject** http://purl.org/dc/elements/1.1/subject	**Materias y palabras clave: El tema del contenido del recurso.** Un tema será expresado como palabras clave, frases clave o códigos de clasificación que describan el tema de un recurso. Se recomienda seleccionar un valor de un vocabulario controlado o un esquema de clasificación formal.
Description (descripción)	**DC. Description** http://purl.org/dc/elements/1.1/description	**Descripción: La descripción del contenido del recurso.** La descripción puede incluir, pero no se limita a: un resumen, tabla de contenidos, referencia a una representación gráfica de contenido o una descripción de texto libre del contenido.
Publisher (editor) y, otras colaboraciones	**DC. Publisher** http://purl.org/dc/elements/1.1/publisher	**Editor: La entidad responsable de hacer que el recurso se encuentre disponible.** Ejemplos de editores son una persona, una organización o un servicio. Típicamente, el nombre de un editor podría usarse para indicar la entidad.
Contributor (otros autores/colaboradores)	**DC. Contributor** http://purl.org/dc/elements/1.1/contributor	**Colaborador. La entidad responsable de hacer colaboraciones al contenido del recurso.** Ejemplos de colaboradores son una persona, una organización o un servicio. Típicamente, el nombre del colaborador podría usarse para indicar la entidad.

Date (fecha)	**DC. Date** http://purl.org/dc/elements/1.1/date	**Fecha: Una fecha asociada con un evento en el ciclo de vida del recurso.** Típicamente, la fecha será asociada con la creación o disponibilidad del recurso. Se recomienda utilizar un valor de datos codificado definido en el documento "Date and Time Formats", http://www.w3.org/TR/NOTE-datetime que sigue la norma ISO 8601 que sigue el formato YYYY-MM-DD.
Type (tipo de recurso)	**DC. Type** http://purl.org/dc/elements/1.1/type	**Tipo: la naturaleza o categoría del contenido del recurso.** El tipo incluye términos que describen las categorías generales, funciones, géneros o niveles de agregación del contenido. Se recomienda seleccionar un valor de un vocabulario controlado (por ejemplo, el *DCMI Vocabulary* -DCMITYPE-http://Dublin Core.org/documents/dcmi-type-vocabulary/). Para describir la manifestación física o digital del recurso, se usa el elemento Formato.
Format (formato)	**DC. Format** http://purl.org/dc/elements/1.1/format	**Formato: la manifestación física o digital del recurso.** El formato puede incluir el tipo de media o dimensiones del recurso. Podría usarse para determinar el *software*, *hardware* u otro equipamiento necesario para ejecutar u operar con el recurso. Ejemplos de las dimensiones son el tamaño y la duración. Se recomienda seleccionar un valor de un vocabulario controlado (por ejemplo, la lista de Internet Media Types (MIME) que define los formatos de medios de ordenador).

Identifier (identificador)	**DC. Identifier** http://purl.org/dc/elements/1.1/identifier	**Identificación: Una referencia no ambigua para el recurso dentro de un contexto dado.** Se recomienda identificar el recurso por medio de una cadena de números de conformidad con un sistema de identificación formal, tal como un URI (que incluye el Uniform Resource Locator -URL, el Digital Object Identifier (DOI) y el International Standard Book Number (ISBN).
Source (fuente)	**DC. Source** http://purl.org/dc/elements/1.1/source	**Fuente**: **Una referencia a un recurso del cual se deriva el recurso actual.** El recurso actual puede derivarse, en todo o en parte, de un recurso fuente. Se recomienda referenciar el recurso por medio de una cadena o número de conformidad con un sistema formal de identificación.
Languaje (lenguaje)	**DC. Language** http://purl.org/dc/elements/1.1/language	**Lengua**: **La lengua del contenido intelectual del recurso.** Se recomienda usar RFC 3066 http://www.ietf.org/rfc/rfc3066.txt en conjunción con la ISO 639 [ISO639] http://www.loc.gov/standards/iso639-2/, que define las etiquetas de dos y tres letras primarias para lenguaje, con subetiquetas opcionales. Ejemplo: "en" u "eng" para Inglés, "akk" para Acadio, y "en-GB" para inglés usado en Reino Unido.
Relation (relación)	**DC. Relation** http://purl.org/dc/elements/1.1/relation	**Relación**: **Una referencia a un recurso relacionado.** Se recomienda referenciar el recurso por medio de una cadena de números de acuerdo con un sistemas de identificación formal.

Coverage (cobertura).	**DC. Coverage** http://purl.org/dc/elements/1.1/coverage	**Cobertura: La extensión o ámbito del contenido del recurso.** La cobertura incluiría la localización espacial (un nombre de lugar o coordenadas geográficas), el período temporal (una etiqueta del período, fecha o rango de datos) o jurisdicción (tal como el nombre de una entidad administrativa). Se recomienda seleccionar un valor de un vocabulario controlado (por ejemplo, del Thesaurus of Geographic Names (TGN) y que, donde sea apropiado, se usen preferentemente los nombres de lugares o períodos de tiempo antes que los identificadores numéricos tales como un conjunto de coordenadas o rangos de datos.
Rights (derechos).	**DC. Rights** http://purl.org/dc/elements/1.1/rights	**Derechos**: **La información sobre los derechos de propiedad y sobre el recurso.** Este elemento podrá contener un estamento de gestión de derechos para el recurso, o referencia a un servicio que provea tal información. La información sobre derechos a menudo corresponde a los derechos de propiedad intelectual, copyright y otros derechos de propiedad.

Tabla 2. Elementos DC (Hipertexto.info 1995)

(Fuente: Hipertexto.info (1995). Metadatos Dublin Core. [online]. Recuperado de: http://www.hipertexto.info/documentos/dublin_core.htm*)*

3 PROTOCOLOS DE INTERCAMBIO DE INFORMACIÓN

Una de las características esenciales de un repositorio digital es compartir o intercambiar información para beneficio de aumentar su visibilidad y por ende la probabilidad de ser consultado. El repositorio digital debe incorporar un mecanismo para que pueda ser accedido por diferentes motores de búsqueda y otras herramientas permitiendo exponer sus metadatos a otros servicios de cosecha y búsqueda de contenidos. Es por esto que cobra vital importancia el papel que toman los protocolos de intercambio de información para realizar estas tareas.

Se sabe que la información es una pieza vital de todo sistema, por lo tanto es importante el poder manejarla en cualquier momento y en cualquier lugar y no solo eso, sino que también debe hacerse de una manera eficiente, rápida y segura. Algunos protocolos de comunicación no solo permiten el intercambio de información, sino que también proporcionan un nivel de seguridad para el envío de nuestros datos.

A continuación se describen algunos de los protocolos más importantes para intercambiar información muchos de los cuales se muestran por (*García, Cárdenas. 2007*)

3.1 HTTP

El protocolo de transferencia de hipertexto (Hyper Text Transfer Protocol) es un protocolo usado para la transferencia de información entre sistemas, de forma clara y rápida. Este protocolo ha sido usado por el World-Wide Web desde 1990.

El funcionamiento de HTTP es mediante el modelo cliente-servidor y se siguen los siguientes pasos:

- El cliente se conecta al servidor.
- El cliente envía una petición.
- El servidor responde a la petición.

3.1.1 Sintaxis y Funcionamiento de HTTP

HTTP se basa en el modelo cliente-servidor. El cliente envía un mensaje al servidor, en donde se incluyen:

- Un comando.
- Un identificador de recurso (URI).
- La versión del protocolo.
- El contenido del mensaje.

El servidor envía un mensaje al cliente, en donde se incluyen:

- La versión del protocolo.
- El código de respuesta.
- El contenido del mensaje.
- La conexión es iniciada por el cliente y cerrada por el servidor después de enviar la respuesta. La versión 1.1 permite conexiones persistentes.
- HTTP por default utiliza el protocolo 80.

3.1.2 Solicitudes

Los mensajes que van del cliente al servidor se llaman solicitudes.

Las solicitudes constan de:

- Una línea de solicitud.
- Una serie de cabeceras.
- Una línea en blanco.
- El cuerpo del mensaje.

La línea de solicitud consta de:

- Un método o comando.
- La URI del recurso solicitado.
- La versión de HTTP a utilizar.

El cuerpo del mensaje es información que se desea enviar al servidor.

3.1.3 Respuestas del HTTP

Los mensajes que van del servidor al cliente se llaman respuestas.

Las respuestas constan de:

- Línea de estado.
- Una serie de cabeceras.
- Una línea en blanco.
- El cuerpo del mensaje.

En el cuerpo del mensaje se incluye la información enviada por el servidor. En caso de respuesta a los métodos GET y POST en el cuerpo se incluye el recurso solicitado.

La línea de estado está formada por:

- La versión del protocolo.
- El código de estado.
- La frase asociada al código.

Los códigos de estado son números de tres dígitos. Se dividen en 5 categorías según su primer dígito.

- 1XX. Informativo (no se utiliza).
- 2XX. Éxito.
- 3XX. Redirección.
- 4XX. Error del cliente.
- 5XX. Error del servidor.

3.1.4 Métodos del HTTP

Los métodos de HTTP son:

OPTIONS: Consultar los métodos asociados a un recurso.
GET: Descargar un recurso.
HEAD: Ver las cabeceras que se envían con un recurso.
POST: Enviar un elemento al servidor.
PUT: Colocar información enviada en la URI identificada.
DELETE: Eliminar la entidad indicada por la URI.
TRACE: Obtiene una réplica del mensaje enviado.

3.2 OAI-PMH

La Open Archives Initiative Protocol for Metadats Harvesting (conocido como el OAI-PMH) proporciona un marco de interoperabilidad independiente de aplicaciones basado en la recopilación de metadatos. (*Openarchives.org. 2002*)
Es el protocolo surgido de OAI (Open Archives Initiative) cuya política es la de compartir información de manera abierta.

Hay dos tipos de roles en el marco OAI-PMH (*véase figura 6*):

- Proveedores de datos de administración de sistemas que soportan el protocolo OAI-PMH como una forma de exponer los metadatos, y
- Los proveedores de servicios utilizan metadatos recolectados a través del protocolo OAI-PMH como base para la creación de servicios de valor agregado. (*Huaroto, L. 2007*)

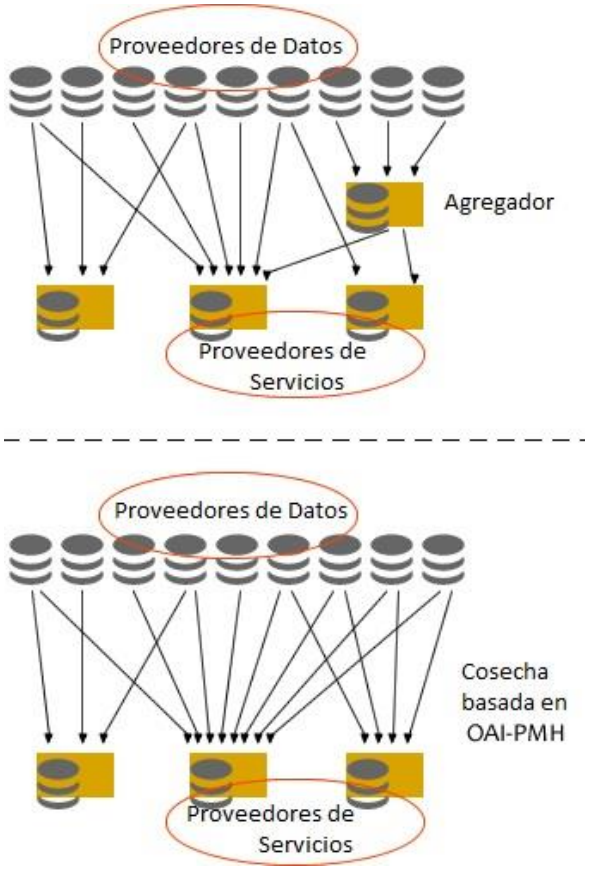


Figura 6. Roles del protocolo OAI-PMH

Se sigue el modelo cliente-servidor, esto es:

- El cliente solicita la información a través del protocolo HTTP, utilizando comandos enviados a través de una URI.
- El servidor contesta utilizando un lenguaje definido en XML para representar Dublin Core o algún otro estándar.

3.2.1 Comandos o Verbos

- getRecord. Regresa un registro dado su identificador. Los argumentos son:
  • identifier. El identificador del registro.
  • metadataPrefix. Especifica el formato de devolución de los resultados.
- identify. Da información acerca del servidor.
- listMetadataFormats. Regresa los formatos disponibles para desplegar los registros.
  El único argumento es:
  • identifier. Este argumento se usa si se desean conocer los formatos para un registro en particular.

- listIdentifiers. Regresa una lista de identificadores de recursos en un rango de fechas dado. Los argumentos son:
  • from. Fecha de inicio.
  • until. Fecha nal.
  • metadataPrex. Especica el formato de devolución de los resultados.
  • resumptionToken. Se utiliza cuando los resultados vienen en páginas, para recorrer estas (es un argumento exclusivo).
- listRecords. Regresa una lista de recursos en un rango de fechas dado. Los argumentos son:
  • from. Fecha de inicio.
  • until. Fecha final.
  • metadataPrefix. Especifica el formato de devolución de los resultados.
  • resumptionToken. Se utiliza cuando los resultados vienen en páginas, para recorrer estas (es un argumento exclusivo).

3.2.2 Ejemplos

- *http://repositalcuaed.unam.mx:8080/oai/request?verb=Identify*
- *http://unibio.unam.mx/oai/request?verb=Identify*
- *http://www.ru.tic.unam.mx:8080/oai/request?verb=ListRecords&from=2002-05-01T14:15:00Z&until=2012-05-01T14:20:00Z&metadataPrefix=oai_dc*
- *http://repositalcuaed.unam.mx:8080/oai/request?verb=ListMetadataFormats*
- *http://repositalcuaed.unam.mx:8080/oai/request?verb=ListMetadataFormats&identifier=oai:132.248.48.117:123456789/651*
- *http://www.ru.tic.unam.mx:8080/oai/request?verb=ListIdentifiers&from=2010-01-01&metadataPrefix=oai_dc*
- *http://unibio.unam.mx/oai/request?verb=ListIdentifiers&from=2010-01-01&metadataPrefix=oai_dc*
- *http://www.ru.tic.unam.mx:8080/oai/request?verb=GetRecord&identifier=oai:ru.tic.unam.mx:DGTIC/59984&metadataPrefix=oai_dc*
- *http://repositalcuaed.unam.mx:8080/oai/request?verb=GetRecord&identifier=oai:132.248.48.117:123456789/651&metadataPrefix=oai_dc*
- *http://arXiv.org/oai2?verb=ListIdentifiers&from=2000-01-01&metadataPrefix=oai_dc*
- *http://arXiv.org/oai2?verb=ListIdentifiers&resumptionToken=391951|10001*

Algunos proveedores de servicio son:

- Greenstone
- DSpace
- SharePoint
- Alfresco

Algunos proveedores de datos son:

- DSpace
- Cybertesis
- E-LIS

3.3 Z39.50

Es una norma establecida para consultar catálogos de bibliotecas a través de Internet.

Éste estándar (ANSI/NISO Z39.50-2003), define un protocolo de aplicación para la búsqueda y recuperación de información en bases de datos.

Hoy en día Z39.50 es quizá la norma más importante en el mundo de las bibliotecas.
Es mantenido por la Biblioteca del Congreso de los Estados Unidos. ANSI/NISO Z39.50, ISO 2350.
También sigue el modelo cliente-servidor, donde el cliente se conecta y extrae información de un servidor, que se encuentra agrupada en bases de datos (*véase figura 7*).
El cliente envía peticiones al servidor y este envía respuestas a ella.
Los pasos que se siguen para la extracción de información son:

- Inicialización.

  Es precursora del trabajo real, en la que se establecen los parámetros básicos de la sesión que se va a iniciar entre el cliente y el servidor. Esta negociación incluye la versión del protocolo, las operaciones que podrán efectuarse, juegos de caracteres, lenguas segmentación y tamaño de la información, etc. Permite asimismo la autenticación del usuario

- Búsqueda.

  Es la funcionalidad más importante del estándar, que permite realizar búsquedas simples o complejas con la misma herramienta a múltiples bases de datos, agilizando la recuperación de información. Los parámetros de búsqueda, en el caso de los registros bibliográficos, están definidos en el set de atributos Bib-1. Las estrategias de búsqueda pueden utilizar operadores booleanos, de proximidad, etc.

- Recuperación.

  Una vez realizada la búsqueda, el cliente solicita al servidor los registros que quiere visualizar, que dependiendo del número solicitado, podrán aparecer segmentados en conjuntos de registros. (*Martínez, Gallo. nd*)

El estándar ofrece otras muchas facilidades y características adicionales, aunque no es necesario que se implanten todas. Algunas de ellas son: controlar el acceso, realizar búsquedas utilizando índices, ordenar la información recuperada, y poder acceder a información sobre el servidor y los servicios que ofrece.

También, el Z39.50 se apoya en lo siguiente:

- ASN.1 (Abstract Syntax Notation One) es una norma creada para el intercambio de estructuras de datos entre aplicaciones.
- BER (Basic Encoding Rules) es el mecanismo utilizado por ASN.1 para identificar y delimitar la información a transferir.
- RPN (Reverse Polish Notation) un lenguaje bajo ASN.1 para describir las expresiones de búsqueda.
- Bib-1. Un conjunto de atributos para apoyar a las búsquedas donde se especifica: el campo, la posición, truncación, completes, etc.
- MARC 21 el formato en el que generalmente se presentan los recursos recuperados. En ocasiones estos se pueden presentar en otros formatos como XML o SUTRS.

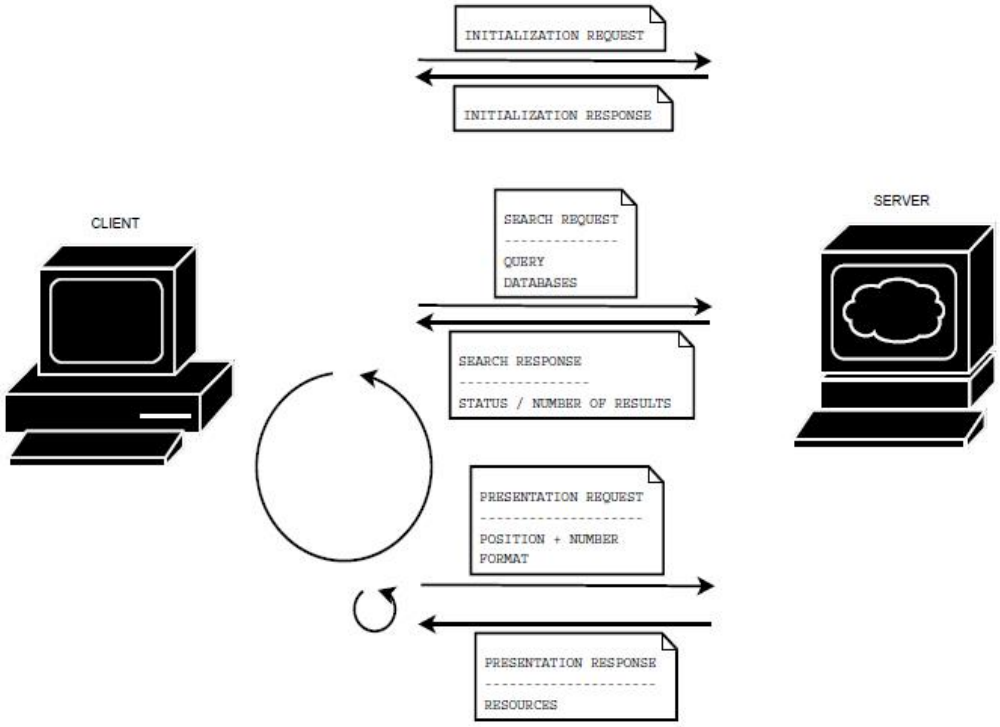


Figura 7. Modelo cliente-servidor del z39.50

3.3.1 Especificación de Z39.50

Z39.50 funciona a base de envío de peticiones por parte del cliente y devolución de respuestas por parte del servidor.

Las peticiones y respuestas corresponden con estructuras de datos definidas mediante ASN.1, utilizando tipos contextuales.

El mecanismo de extracción de resultados en Z39.50 es el siguiente:

El cliente envía una petición de inicialización.
El servidor contesta con una respuesta de inicialización.
El cliente envía una petición de búsqueda al servidor.
El servidor contesta con una respuesta de búsqueda.
El cliente envía una petición de presentación al servidor.
El servidor contesta con una respuesta de presentación.

La petición de inicialización (etiqueta 20) consta de:

Versión del protocolo (etiqueta 3).
Opciones (etiqueta 4).
Tamaño preferido del mensaje (etiqueta 5).
El tamaño máximo de registro (etiqueta 6).
El identificador de autenticación (etiqueta 7).
El identificador de implantación (etiqueta 110).
El nombre de implantación (etiqueta 111).
La versión de implantación (etiqueta 2.0.21).

La respuesta de inicialización (etiqueta 21) consta de:

El identificador de referencia (etiqueta 2).
Versión del protocolo (etiqueta 3).
Opciones (etiqueta 4).
Tamaño preferido del mensaje (etiqueta 5).
El tamaño máximo de registro (etiqueta 6).
La bandera de resultado (FF verdadero) (etiqueta 12).
El identificador de implantación (etiqueta 110).
El nombre de implantación (etiqueta 111).
La versión de implantación (etiqueta 2.0.21).

La petición de búsqueda (etiqueta 22) consta de:

- Nombre del conjunto de resultados (etiqueta 17).
- Nombres de las bases de datos (etiqueta 18).
- Consulta (etiqueta 21), utilizando RPN.

La respuesta de búsqueda (etiqueta 23) consta de:

- El identificador de referencia (etiqueta 2).
- Contador de resultados (etiqueta 23).
- Número de registro regresado (etiqueta 24).
- Posición siguiente en el conjunto de resultados (etiqueta 25).
- La bandera de resultado (FF verdadero) (etiqueta 22).
- Estado de presentación (etiqueta 27).

La petición de presentación (etiqueta 24) consta de:

- Identificador del conjunto de resultados (etiqueta 31).
- Punto de inicio en el conjunto de resultados (etiqueta 30).
- Número de resultados pedidos (etiqueta 24).
- Sintaxis preferida para los registros (etiqueta 104).

La respuesta de presentación (etiqueta 25) consta de:

- El identificador de referencia (etiqueta 2).
- Número de registros regresados (etiqueta 24).
- Posición siguiente en el conjunto de resultados (etiqueta 25).
- Estado de presentación (etiqueta 27).
- Resultados o diagnóstico (etiqueta 28).

3.3.2 Estructura RPN

Se utiliza RPN (Reverse Polish Notation) para construir expresiones de búsqueda, mediante ASN.1.

Los componentes de una expresión de búsqueda son:

- Operandos (etiqueta 102)
  - Atributos (etiqueta 44)
  - Término (etiqueta 45)
- Operadores (etiqueta 46)
  - And (valor 0).
  - Or (valor 1)
  - Not (valor 2)

3.3.3 BIB-1

Atributos de BIB-1:

1. Atributos de uso.
2. Atributos de relación.
3. Atributos de posición.
4. Atributos de estructura.
5. Atributos de truncado.
6. Atributos de completes.

3.3.4 Otros acuerdos

El estándar también especifica el manejo y ordenación de los juegos de resultados, presentación de índices, apertura y cierre de sesiones y extensiones a otros servicios no definidos en el estándar mismo.

Además, z39.50 define otros aspectos, como:

• Lenguaje de búsquedas basado en juegos de atributos.
• Sintaxis de resultados aceptables (MARC, GRS-1).
•Lenguaje de construcción de resultados para su trasferencia.
•Facultad para que el servidor pueda explicar su estructura interna y sus capacidades.

La versión 3 (1995) de la norma z39.50 permite establecer un sistema de búsqueda muy potente, que puede incluir:

- Todos los operadores booleanos (que no implantan en la actualidad la mayoría de los clientes).
- Operadores de comparación de fechas (greater than, equal to...).
- Operadores de proximidad.
- Diversas opciones para realizar el truncamiento.
- Búsquedas completas (part of field, complete field...)

Además, existen otros rasgos adicionales que ofrecen múltiples posibilidades:

- Autentificación: esto permite que el servidor-Z pueda controlar quién accede a las bases de datos.
- Control de los recursos y de los accesos (cuentas).
- Opción “explain”, que permite obtener información sobre bases de datos remotas, servicios disponibles...
- “Browsing” del índice.
- Definir el formato de los registros.

3.3.5 Servicios Extendidos

La versión 3 define también el uso de la norma para implantar lo que denomina como “servicios extendidos”. A pesar de no estar incluido dentro de la norma, z39.50 incluye los siguientes servicios para facilitar el control:

- Almacenar resultados.
- Almacenar una query.
- Definir un esquema de búsqueda.
- Solicitar un ejemplar.
- Actualizar la base de datos.
- Crear un fichero de exportación

3.4 SOAP

SOAP (Simple Object Access Protocol). Es un protocolo que se basa en XML para permitir el intercambio de información entre aplicaciones a través de HTTP.

SOAP funciona a través del envío de mensajes entre programas, estos mensajes contienen llamadas y respuestas como se ve en la figura 8.

Un mensaje en SOAP es un documento XML que incluye los siguientes elementos:

- Envelope. Identifica al documento XML como un mensaje SOAP.
- Header. Un encabezado opcional.
- Body. Contiene las llamadas y/o respuestas.
- Fault. Contiene información de errores ocurridos.

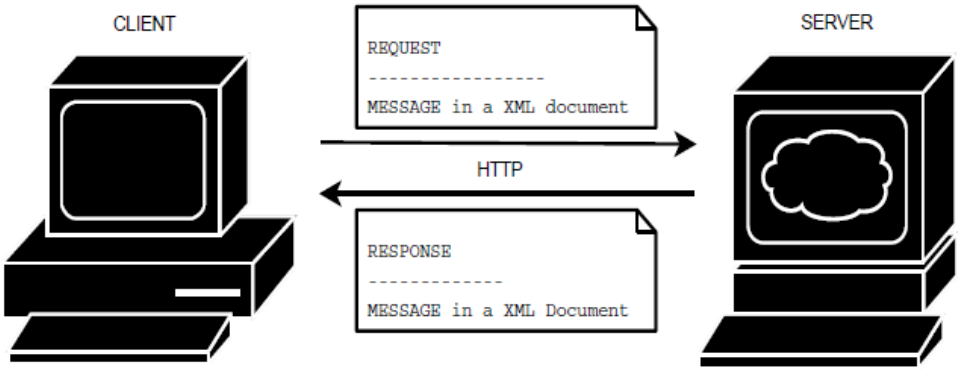


Figura 8. Modelo cliente-servidor SOAP

3.4.1 Objetivos primordiales de SOAP

a) Establecer un protocolo estándar de invocación de servicios remotos, basado en protocolos estándares de Internet: HTTP (Protocolo de transporte de Hipertexto) para la transmisión y XML (lenguaje de marcado extensible) para la codificación de datos.

b) Independencia de plataforma, lenguaje de desarrollo e implantación (modelo de objetos).

El protocolo de comunicación HTTP es el empleado intrínsecamente para la conexión sobre Internet. Garantiza que cualquier cliente con un navegador estándar pueda conectarse con un servidor remoto. La transmisión de datos se empaqueta con XML, que se ha convertido en el estándar del intercambio de datos, salvando las incompatibilidades entre otros protocolos, tales como el NDR (Network Data Representation) o el CDR (Common Data Representation).

Por otra parte, los servidores Web pueden procesar las peticiones de usuario, empleando las tecnologías de Servlets, paginas ASP (Active Server Pages) o JSP (Java Server Pages), o un servidor de aplicaciones, invocando objetos de tipos CORBA, COM o EJB.

3.4.2 Funcionamiento de SOAP

La especificación SOAP menciona que las aplicaciones deben ser independientes del lenguaje de desarrollo, por lo que las aplicaciones cliente y servidor pueden estar escritas con HTML, DHTML, Java, Visual Basic u otras herramientas y lenguajes disponibles. Lo importante es tener alguna implantación de SOAP (dependiendo de la herramienta de desarrollo elegida) y enlazar sus librerías con la aplicación. Aunque esto no es estrictamente necesario, es preferible trabajar usando dichas librerías, con el fin de no reescribir un código ya probado. (*Botello, Castillo. 2002*)

Las peticiones con el uso del protocolo HTTP emplean el comando POST para transmitir información entre el cliente y el servidor.

Por otra parte el término Object en el nombre significa que se adhiere al paradigma de la programación orientada a objetos.

SOAP es un marco extensible y descentralizado que permite trabajar sobre múltiples pilas de protocolos de redes informáticas. Los procedimientos de llamadas remotas pueden ser modelados en la forma de varios mensajes SOAP interactuando entre sí.

Estos mensajes constan de 3 secciones: envelope, header y body (*véase figura 9*).

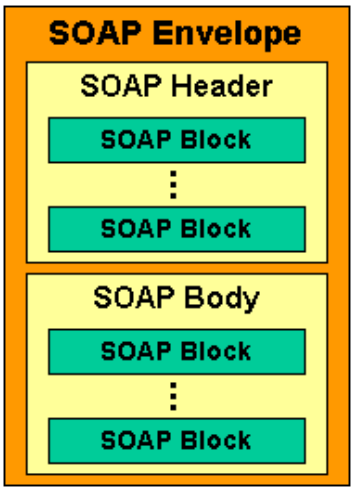


Figura 9. Modelo de encapsulamiento ilustrando las partes de un mensaje SOAP

Donde:

- envelope (envoltura): Es el elemento raíz del mensaje para describir su contenido y la forma de procesarlo.
- header (encabezado): Es la información de identificación del contenido. Un grupo de reglas de codificación para expresar las instancias de tipos de datos definidos por la aplicación.
- body (cuerpo): Es el contenido del mensaje. Una convención para representar las llamadas y las respuestas a procedimientos remotos (*W3,2001*).

3.5 SRW

SRW (Search/Retrieve Web Service) es una variación de SRU (Search/Retrieve via URL), ambos son estándares para realizar consultas en Internet.

Se basa en los siguientes estándares:
CQL (Common Query Language) para representar consultas.
SOAP como protocolo base de comunicación.

Es mantenido por la biblioteca del Congreso de Estados Unidos.

3.5.1 Funcionamiento de SRW

SRW sigue el modelo cliente-servidor (*véase figura 10*) y funciona de la siguiente manera:

El cliente envía una petición al servidor.
El cliente regresa la respuesta a dicha petición.

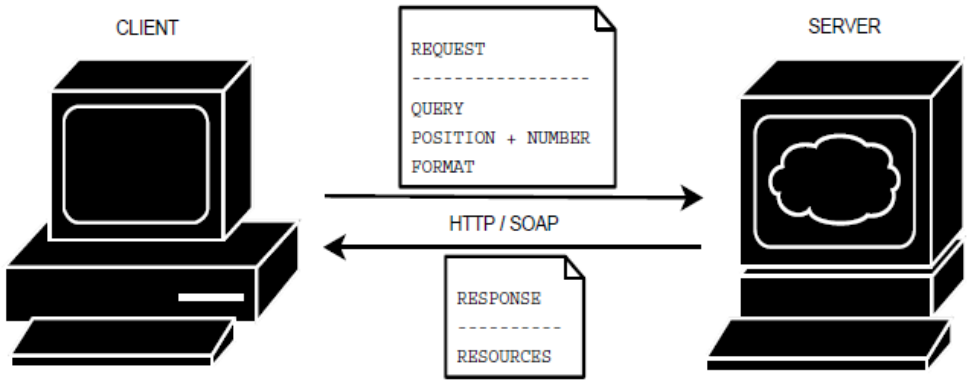


Figura 10. Funcionamiento de SRW

3.5.2 Parámetros de Petición

- version. (Obligatorio). La versión del protocolo en la petición.
- query. (Obligatorio). La expresión de consulta en CQL.
- startRecord. (Opcional). La posición del primer registro a ser recuperado en la secuencia de resultados.
- maximumRecords. (Opcional). El número de registros a ser devueltos.
- recordPacking. (Opcional). Determina como se presentan los registros (string, xml).
- recordSchema. (Opcional) El esquema solicitado para los registros a recuperar.

- recordXPath. (Opcional). Una expresión en Xpath para aplicar a los registros antes de regresarlos.
- resultSetTTL. (Opcional). El número de segundos a esperar por una respuesta.
- sortKeys. (Opcional). Contiene una secuencia de claves de búsqueda para aplicar a los resultados.
- extraRequestData. (Opcional). Contiene información extra como perfiles.

3.5.3 Parámetros de Respuesta

- version. (Obligatorio). La versión del protocolo en la respuesta.
- numberOfRecords. (Obligatorio). El número de registros que corresponden a la consulta (0 búsqueda fallida).
- resultSetId. (Opcional). El identificador del conjunto de resultados creado por la búsqueda.
- resultSetIdleTime. (Opcional). El número de segundos para borrar el conjunto de resultados.
- records. (Opcional). La secuencia de registros.
- nextRecordPosition. (Opcional). La siguiente posición en el conjunto de resultados.
- diagnostics. (Opcional). Secuencia de diagnósticos.
- extraResponseData. (Opcional). Información adicional.

3.6 Otros estándares involucrados en la recuperación de información

3.6.1 OpenURL

OpenURL es una norma que permite entrelazar referencias Bibliográficas.

En OpenURL se estandarizan las direcciones URL para que empleen los metadatos que posee el usuario.

Se necesita de un servidor (resolver) que interprete la URL estandarizada dada por el cliente.

Open URL hace uso de HTTP pero no se especifica la manera en que se despliegan los resultados.

3.6.2 DOI

El DOI (Digital Object Identifier) es un número de control creado para identificar unívocamente objetos digitales.

Surgió en 1997 como una iniciativa del Comité de Tecnologías de la Asociación Americana de editores y desde 1998 es un proyecto gestionado por la "International DOI Foundation".

Los objetivos planteados con la creación de este número son:

- Proporcionar un marco para la gestión de la propiedad intelectual.
- Favorecer la comunicación entre clientes (lectores) y editores.
- Facilitar el comercio electrónico.
- Posibilitar la gestión del copyright de forma automatizada.

La tecnología DOI se basa en el programa "Handle System" desarrollado por la "Corporation for National Research Initiatives" a partir de 1997.

El DOI básicamente es un código alfanumérico (también conocido como nombre DOI) que incluye dos partes: prefijo y sufijo.

- El prefijo a su vez consta de dos componentes: la primera componente identifica la secuencia como DOI y la segunda identifica la editorial.
- El sufijo identifica el objeto digital (por ejemplo el ISSN).

Actualmente existen más de 25 millones de DOI's registrados en las distintas agencias.

La utilidad del DOI también se deriva de una serie de servicios adicionales asociados al propio registro.

Recientemente se ha adoptado el DOI como base para el intercambio de información en el marco del proyecto "Crossref".

4 PRESERVACIÓN DIGITAL

4.1 Introducción a la preservación digital

Las colecciones digitales crecen a un ritmo acelerado, como ha sucedido durante los últimos veinte años. Este crecimiento sostenido y, hasta ahora incontrolado, plantea la necesidad de procedimientos que garanticen no sólo la permanencia de las colecciones, sino también que sean consultables y recuperables, independientemente de los cambios tecnológicos. (*Lara, Pacheco. 2008*)

4.2 Definición de preservación digital

El proceso de preservación digital amerita la siguiente reflexión:

¿Preservar es lo mismo que transferir documentos originales a un formato digital, o se refiere a preservar los documentos digitales ya creados?

Ambos enfoques son válidos, existen muchos proyectos de preservación de originales por métodos digitales. Muchas bibliotecas están involucradas en proyectos de digitalización de fondos históricos para mejorar el acceso y, además, contribuir a la preservación del original, ya que el uso de su copia virtual le protege de los efectos nocivos de la manipulación física (*Keefer,A. 2003*) (*Bia, A. 2002*). En algunos medios analógicos, tal como las cintas magnéticas, la digitalización ayuda a proteger la calidad de la información (videos por ejemplo) de la degradación natural que sufre el medio en el transcurso del tiempo. El enfoque de preservación de los propios materiales digitales se da debido a la gran fragilidad de los medios de almacenamiento de información digital aunado a los avances rápidos de la tecnología y a la rápida obsolescencia de medios de almacenamiento, hardware y software.

La idea de preservar los documentos digitales surge por la fragilidad de los soportes de almacenamiento de la propia información digital, sin contar los rápidos avances de la tecnología y la continua obsolescencia de los soportes de almacenamiento, el hardware y el software.

Según (*Jones, M. 2001*), la preservación digital se refiere a una serie de actividades necesarias y muy bien administradas para asegurar el acceso continuo a los materiales digitales, por el periodo que sea necesario. Se refiere a todas las acciones requeridas para mantener el acceso a los materiales digitales aún después de que se presenten fallas en los medios de almacenamiento o haya

cambios tecnológicos. La preservación se clasifica en tres grupos de acuerdo al tiempo:

- **Preservación de duración larga:** Acceso continuo a los materiales digitales o por lo menos a la información contenida en éstos indefinidamente.
- **Preservación de duración media:** Acceso continuo a los materiales digitales aún después de los cambios tecnológicos realizados en un periodo definido de tiempo pero no indefinidamente.
- **Preservación de duración corta:** Acceso a los materiales digitales ya sea por un periodo de tiempo definido o que su uso sea calculado en un periodo de tiempo menor a los cambios tecnológicos.

4.3 Definición de respaldo

El respaldo, también conocido como copia de seguridad, se refiere a la existencia de una réplica de los datos o la información de un sistema, para que éste pueda ser restaurado en caso de fallas o desastres. En este sentido un respaldo es utilizado como un plan de contingencia, para restaurar un equipo de cómputo a un estado operacional luego de un desastre, o bien, para recuperar datos o información que se hayan borrado o corrompido por cualquier causa.

4.4 Diferencias entre preservación y respaldo

La preservación digital es diferente de las copias seguridad. Lo que se guarda como copia de seguridad en una biblioteca digital son, básicamente, dos cosas: por un lado la información publicada en el servidor (recursos digitales más información de catálogo) y, por otro lado, los recursos digitales en proceso de edición. La preservación digital sin embargo, no se ocupa de respaldar ni los datos del servidor ni el material de trabajo diario, sino de salvaguardar los recursos digitales que necesitaremos en el futuro (*Bia, A. 2002*). Debido a la limitante en el ancho de banda de red, de muchos usuarios de bibliotecas digitales, comúnmente la información publicada en el servidor está comprimida o sacrifica su calidad para reducir su tamaño y pueda descargarse fácilmente. La información digital, que se desea preservar, debe de ser de la máxima calidad posible para usos futuros. Tal como se indica en (*McGray, A.T. 2001*) debe realizarse una separación entre el material para archivo y los derivados para acceso público. Este modelo de biblioteca digital incluye una versión maestra de la biblioteca digital con los recursos de alta calidad (los que se preservan) y una biblioteca de acceso público con formatos generados automáticamente a partir de la primera.

Si bien las copias de seguridad, al igual que las de preservación, se basan en la redundancia de la información mediante grabaciones periódicas, ni la forma de organizar estas grabaciones ni los tiempos son los mismos. Las copias de seguridad pueden seguir diversos métodos conocidos: copia integral, copia incremental o copias rotativas, por ejemplo, y la periodicidad generalmente es alta (diaria o semanal). En el caso de la copias de preservación, por el contrario, el método suele ser la grabación integral del material una vez y el copiado del mismo una vez al año o cada año y medio en otro soporte nuevo (rejuvenecimiento) (*Bia, A. 2002*).

4.5 Problemáticas en la preservación digital

En un estudio realizado en (*Preserving our digital heritage, October 2002*) se detectaron un conjunto de problemáticas de preservación digital describiendo un grupo general e ilustrando algunas particulares de acuerdo a los siguientes recursos electrónicos: libros, revistas, grabaciones de sonidos, televisión digital, video y páginas WEB. Estas problemáticas, junto para la preservación digital, junto con algunas adiciones, se resumen a continuación:

- Nuevos enfoques de seleccionar y catalogar
- Multiplicidad de formatos
- Cambios rápidos en la tecnología
- Obsolescencia de hardware y software
- Problemas legales, sociales y económicos

4.5.1 Libros electrónicos

- Diversidad de iniciativas de estándares
- Bajo desarrollo en precauciones de seguridad en el mercado
- Dispositivos de hardware y software propietarios

4.5.2 Revistas electrónicas

- Provistas por ligas a proveedores
- Contienen artículos repletos con citas a otros recursos secundarios en línea o ligas que probablemente no se preservan
- ¿Para preservar un artículo hay que preservar todos sus enlaces? ¿tenemos derecho de hacerlo?

4.5.3 Grabaciones de sonidos

- Migración de sistemas analógicos a digitales
- Dependencias de máquinas y medios
- Obsolescencia de medios
- Sistemas de almacenamiento masivo

4.5.4 Televisión digital y video

- Migración de sistemas analógicos a digitales
- Dependencias de máquinas y medios
- Obsolescencia de medios
- Demandan sistemas de gran escala de almacenamiento

4.5.5 WEB

- Mortalidad de enlaces demasiado alta
- Contienen enlaces a otros recursos en línea, de los cuales, algunos probablemente no se preservan
- ¿Para preservar un documento WEB hay que preservar todos sus enlaces? ¿tenemos derecho de hacerlo?
- ¿Cómo definir los límites en los enlaces de un servidor web?

4.6 Respaldo y recuperación

Los sistemas de cómputo que contienen las colecciones digitales, están expuestos a riesgos latentes. Pueden verse interrumpidos en su servicio, debido a alteraciones en la electricidad, el hardware, el software y la red, así como fallas humanas, desastres naturales y ataques informáticos como virus y sabotaje, entre otras.
El riesgo a que están expuestos nuestros sistemas informáticos es inminente, por tal motivo es necesario contar con un plan de contingencia adecuado que garantice la recuperación de la información así como la disponibilidad del sistema informático que la gestiona. De acuerdo con (*Hernández, I. 2005*) para la elaboración del plan de contingencia es necesario lo siguiente:

- Identificar y priorizar los procesos y los recursos indispensables;
- Analizar el riesgo y el impacto por la pérdida de la información;
- Evaluar recomendaciones de protección;
- Contar con estrategias y alternativas de recuperación;
- Establecer los equipos de trabajo y las funciones de cada persona;
- Ejecutar simulacros del plan de contingencia;

- Elaborar un manual de contingencia, y
- Retroalimentar el plan.

Para llevar a cabo un plan de contingencia, es recomendable realizar algunas de las siguientes actividades:

- Seleccionar el medio de almacenamiento secundario;
- Determinar la frecuencia de realización de copias de seguridad;
- Determinar el volumen de la información a respaldar, y
- Determinar días y horario en que deben realizarse los respaldos.

Además de permitir la identificación de la mejor manera de recuperar la información en caso de desastre, una estrategia de recuperación es una guía para el desarrollo de los procedimientos mismos de recuperación.

4.6.1 Respaldo tradicional

El respaldo tradicional consiste en copiar los datos o la información de un sistema a un medio de almacenamiento secundario, como cinta, CD y DVD, entre otros, con el fin de que pueda ser restaurado en caso de fallas o desastres. Su periodicidad puede ser diaria, semanal o mensual y difícilmente menor a un día (Ortíz, Dante. 2010). Para realizar las copias los métodos a seguir pueden ser los siguientes:

- *Copiar sólo los datos*. No proporciona las facilidades para recuperar el entorno operacional que proporcionan los programas de aplicación para acceder a los mismos.
- *Copia completa*. Incluye una copia de datos y programas que permite restaurar el sistema hasta el momento anterior a la copia.
- *Copia incremental*. Solamente se almacenan las modificaciones realizadas después de la última copia de seguridad. Debe mantenerse la copia original para restaurar posteriormente el resto de las copias.
- *Copia diferencial*. Es similar a la incremental, pero en lugar de copiar las modificaciones, son almacenados los archivos completos que han sido modificados. También se necesita la copia original.

4.6.2 Respaldo con tecnología RAID

En el mejor de los casos, el sistema de respaldo tradicional se aplica todos los días, comúnmente por la noche, cuando disminuye la carga de trabajo del servidor. Esto significa que si se presenta un incidente en el transcurso del día o, en las circunstancias más adversas, por la tarde, no sería posible recuperar el trabajo realizado. Para muchas empresas esto puede representar grandes pérdidas financieras. Para muchas empresas esto puede representar grandes pérdidas financieras. En el caso de los bancos, por ejemplo, no pueden perder las transacciones realizadas a lo largo del día. Para este tipo de contingencias la

solución tecnológica es el uso del RAID (*Redundant Array of Inexpensive Disks* o Conjunto redundante de discos baratos y, actualmente, *Redundant Array of Independent Disks* o Conjunto redundante de discos independientes).

En informática, el acrónimo RAID se refiere a un sistema de almacenamiento en el que se usan múltiples discos duros, entre los que son distribuidos o replicados los datos. Dependiendo de su configuración, a la que suele denominarse “nivel”, los beneficios de un RAID con respecto a un único disco son:

- Mayor integridad.
- Tolerancia a fallos.
- Rendimiento y capacidad.

En sus orígenes, la principal ventaja de RAID radicaba en su capacidad de combinar varios dispositivos de bajo costo con una tecnología más antigua, para dar como resultado un conjunto que ofrecía mayor capacidad, fiabilidad, velocidad, o una combinación de éstas, que un solo dispositivo de última generación y costo mayor.

En el nivel más simple, RAID combina múltiples discos en una sola unidad lógica: en lugar de identificar diferentes discos, el sistema operativo sólo reconoce uno. Así, el RAID agrupa dos o más discos duros, ofreciendo una forma más avanzada de respaldo, puesto que:

- Es posible mantener copias en línea (redundancia).
- Agiliza las operaciones del sistema, sobre todo en bases de datos.
- El sistema es capaz de recuperar información, sin la intervención de un administrador.

Hablar del nivel o la configuración del RAID, es referirse a la arquitectura que determina la redundancia y cómo están distribuidos los datos a través de los discos duros del arreglo. Existen varias configuraciones del RAID, sin embargo, los cuatro tipos que prevalecen en muchas arquitecturas son RAID-0, RAID-1, RAID-3 y RAID-5.
Cabe aclarar que para la implantación de la tecnología del RAID se requiere un presupuesto mayor. Por este motivo es una tarea importante de las instituciones analizar y evaluar, en función de sus recursos financieros y necesidades, la tecnología a utilizar.

4.7 Estrategias para la preservación digital

Para garantizar la preservación digital en el corto, mediano y largo plazos, dependiendo del tiempo de vida de un documento, existen diferentes estrategias que pueden aplicarse.

4.7.1 Preservación de la tecnología

Para visualizar y editar un contenido digital es recomendable preservar el ambiente tecnológico, incluidos el software y el hardware: sistemas operativos, programas de visualización y periféricos de lectura y escritura de medios de almacenamiento secundario, entre otros.

4.7.2 Migración

La migración es la transferencia o adaptación del contenido digital de una generación de hardware y software a otra, superando la obsolescencia tecnológica, aunque se tiene la desventaja de sufrir pérdidas en la información tras migraciones sucesivas.

Figura 11. Migración de una generación a otra de hardware y software

4.7.3 Reformateo

El reformateo se refiere a cambiar el contenido digital de un formato a otro.

Figura 12. Reformateo de un contenido digital

4.7.4 Refrescado o rejuvenecimiento

Refrescar un contenido digital significa copiarlo de un medio de almacenamiento a otro nuevo del mismo tipo o bien, escribir, cada determinado tiempo, un contenido digital en un medio nuevo, para evitar que el contenido se pierda por la degradación natural que conlleva el transcurso del tiempo.

Figura 13. Migración de una generación a otra de hardware y software

4.7.5 Emulación

Según (Waugh, A. 2000), la emulación permite que el software original sea usado sin necesidad de que el sistema original que lo ejecutaba siga existiendo. La emulación obliga a preservar una cantidad importante de información. Una solución de emulación por hardware, por ejemplo, implica la preservación del emulador, el sistema operativo, la aplicación y los datos.

Un ejemplo de emulación es la ejecución de un sistema operativo dentro de otro.

Figura 14. Emulación de software (ejecución de Linux dentro de Windows)

4.7.6 Replicación

La replicación es la generación y el mantenimiento de una o más copias de un mismo contenido digital.

Figura 15. Replicación de un contenido digital

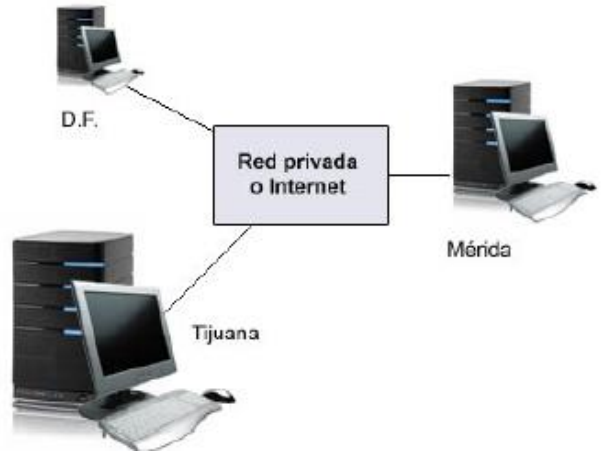


Figura 16. Replicación de grandes contenidos digitales

4.7.7 Estandarización

La estandarización se refiere a la utilización de un formato estándar para la representación de un documento digital, lo que garantiza un mejor soporte de herramientas para administrar la colección digital, una mayor duración del formato y una mejor migración ante los cambios tecnológicos.

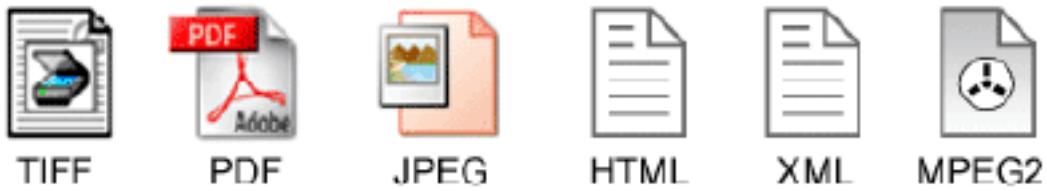


Figura 17. Formatos estándares para representar documentos digitales

4.7.8 Encapsulado

El encapsulado es el empaquetamiento de la información que se desea preservar, junto con un diccionario de datos o metadatos descriptivos, mantenidos en una única localización. Además, incorpora otros factores clave para la preservación de larga duración:

a) auto documentación o la capacidad de entender y decodificar la información preservada sin hacer referencia a información externa;
b) auto suficiencia o minimización de dependencias con respecto a sistemas, datos o información;
c) documentación de contenido o habilidad para que un futuro usuario encuentre o implante el software para visualizar la información preservada, y
d) preservación de organización o habilidad para almacenar la información que permita a una organización el uso eficiente de la información preservada.

Figura 18. Paquete de información

4.7.9 Autenticidad

La autenticidad se refiere al aseguramiento de la integridad de una información digital. Existen muchas causas por las cuales se puede corromper: virus, negligencias, fallas de los medios de almacenamiento, ataques informáticos, etcétera. Para asegurar la autenticidad se propone utilizar huellas, firmas y certificados digitales sobre la información digitalizada.

Ejemplo: La función matemática H genera para el documento digital D una huella digital h(d)=879d8a206e718d8e651a0df1e42ab7007f412a82.

La huella digital es única para cada documento, lo que quiere decir que si dos documentos tienen la misma huella digital, entonces se trata del mismo documento. El proceso de la firma digital para ofrecer autenticidad, es similar al de la firma autógrafa. Un certificado digital es un documento electrónico que demuestra identidad en transacciones electrónicas, validando que una firma digital pertenezca a una entidad identificada. Una autoridad certificadora es el equivalente a un notario.

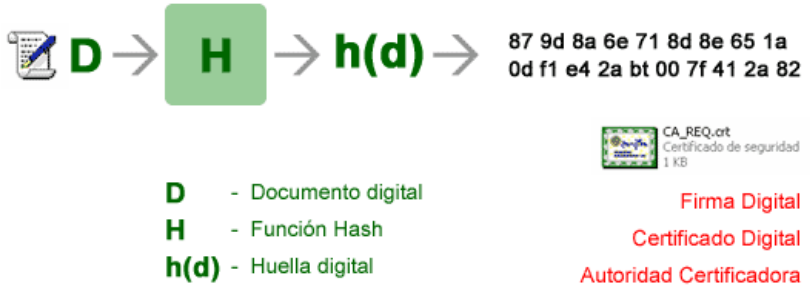


Figura 19. Autenticidad de la información digital

4.7.10 Arqueología Digital

La arqueología digital es un proceso para la recuperación de información, a partir de medios dañados o antiguos de almacenamiento digital.

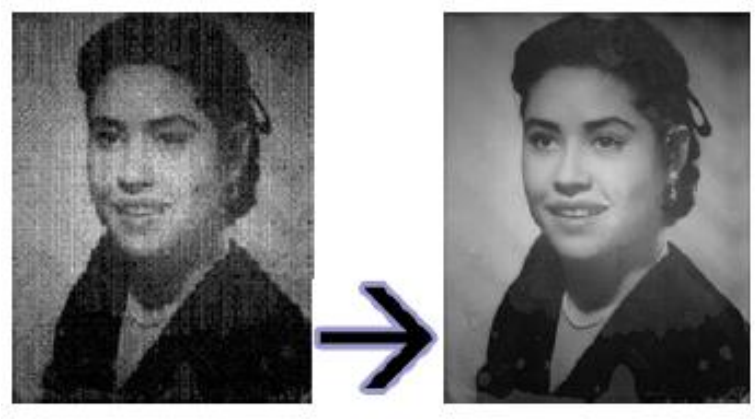


Figura 20. Restauración de imágenes

4.7.11 Cuidado Duradero

El cuidado duradero debe ser visto como una estrategia continua para asegurar que los documentos digitales se encuentren en óptimas condiciones. En el cuidado de una colección los archivos deben almacenarse en medios y ubicaciones no sólo seguros, sino también confiables. Además, deben manipularse con base en las pautas de aceptación internacional, orientadas a optimizar su expectativa y la calidad de duración.

Figura 21. Cuidado duradero

4.8 Modelo de referencia OAIS

El modelo de referencia OAIS (Open Archival Information System) está enfocado a la preservación a largo plazo de la información en formato digital como una manera de garantizar el acceso a ella en el futuro. Consiste básicamente en un modelo lógico sobre la forma como los documentos digitales deben ser preparados, enviados a un archivo, almacenados durante periodos largos, conservados y recuperados.

El modelo de referencia OAIS se ha convertido en el más reconocido para la preservación de información digital. Identifica las responsabilidades y los componentes de un sistema para archivar documentos digitales, incluyendo:

- Las funciones de las personas y las instituciones que interactúan con un documento digital: productor, administrador y consumidor.
- Los objetos digitales o documentos manejados por OAIS, denominados paquetes de información, y
- Seis funciones de alto nivel del modelo: Ingesta, Administración de datos, Almacén de archivos, Acceso, Planeación de la preservación y Administración, que representan treinta y tres funciones de nivel bajo.

4.9 Entidades de OAIS

El diagrama OAIS (*véase figura 22*) ilustra las relaciones entre las funciones. En él los rectángulos identifican grupos de funciones relacionadas. En la práctica no es necesario que las funciones estén en el mismo servidor o en la misma organización. Los grupos pueden estar separados y sus funciones distribuidas en muchas configuraciones.

En el exterior de OAIS se encuentran los productores, los consumidores y los administradores:

- *Productor*. Es la persona que proporciona al sistema la información que va a ser preservada.
- *Administrador*. Es la persona que define las políticas de administración y el control de la administración de OAIS sólo una de sus responsabilidades. No está involucrado en las operaciones diarias del archivo, pues éstas son responsabilidad de la entidad funcional *Administración.*
- *Consumidor*. Interactúa con los servicios de OAIS para encontrar y obtener la información preservada de su interés.

Figura 22. Entidades funcionales de OAIS

El modelo de referencia OAIS está compuesto por seis entidades funcionales y sus interfaces relacionadas. En la figura 22, que muestra el modelo, sólo se presentan los flujos de información más importantes. Las líneas que conectan las entidades, identifican las rutas de información, sobre las cuales ésta fluye en ambas direcciones. Las líneas discontinuas se utilizan para evitar confusión.

Así, el modelo de referencia OAIS, está conformado por:

Ingesta. Esta entidad proporciona los servicios y las funciones para aceptar los Paquetes de Información Sometida (PISs) de los productores o los elementos internos bajo el control de la Administración. Además, prepara el contenido para el manejo y almacenamiento en el archivo.

Las funciones de Ingesta incluyen:

a) La recepción de PISs. aseguran su calidad y generan el Paquete de Información de Archivado (PIA); se encargan también de que cumpla con los estándares de documentación y el formateo de datos.

b) Extracción de información descriptiva de los PIAs para su inclusión en la base de datos del archivo.
c) Coordinación de actualizaciones en Almacén del archivo y Administración de datos.

Almacén del archivo. Proporciona los servicios y las funciones para el almacenamiento, mantenimiento y recuperación de PIAs. Sus funciones incluyen:

a) La recepción de PIAs de Ingesta.
b) Agregado de PIAs para el almacenamiento permanente.
c) Administración de una jerarquía de almacenamiento.
d) Actualización de los medios sobre los cuales los contenedores de los archivos son almacenados.
e) Verificación de errores, brindando capacidades para la recuperación de desastres.
f) Proporcionar PIAs para satisfacer las órdenes generadas por los consumidores.

Administración de datos. Brinda los servicios y las funciones para poblar, mantener y acceder a la información descriptiva, la cual identifica y documenta contenedores de archivos y datos administrativos para el manejo de un archivo.

En sus funciones se incluyen:

a) Administración de la base de datos del archivo, con lo que mantiene las definiciones del esquema, así como vistas e integridad referencial.
b) Ejecución de actualizaciones de la base de datos y carga de información descriptiva nueva o datos administrativos del archivo.
c) Ejecución de consultas sobre datos para la administración de los mismos.
d) Generación de conjuntos de resultados.
e) Generación de reportes.

Administración. Esta entidad proporciona los servicios y las funciones para la operación global del sistema de archivo. Las funciones de administración incluyen:

a) La solicitud y negociación de los acuerdos de sometimiento con los productores.
b) La auditoría de los sometimientos, para asegurar que cumplan con los estándares de archivo.
c) Mantenimiento de la administración de la configuración del software y el hardware del sistema.
d) Proporcionar funciones de ingeniería del sistema para el monitoreo y el mejoramiento de las operaciones del archivo, inventario, reportes y migración/actualización del contenido de un archivo.

Finalmente, la Administración es la responsable de establecer y mantener las políticas y los estándares del archivo, brindando soporte a los usuarios y habilitando las solicitudes almacenadas.

Planeación de preservación. Además de proporcionar los servicios y las funciones para el monitoreo del ambiente de OAIS, esta entidad brinda recomendaciones para asegurar que la información almacenada en el sistema de archivado (es decir, el sistema que permite archivar los documentos digitales) permanezca disponible para la comunidad de usuarios durante un tiempo muy prolongado, incluso si el ambiente original de computación se vuelve obsoleto.

Las funciones de esta entidad abarcan:

a) Evaluación del contenido y recomendaciones periódicas de actualización de información de un archivo para migrar los contenedores actuales de los archivos.
b) Emisión de recomendaciones sobre políticas y estándares de archivo.
c) Monitoreo de cambios en el ambiente tecnológico y en los requerimientos de servicios de los usuarios
d) Constitución de una base de conocimientos de la comunidad de usuarios.

En la planeación de la preservación también son diseñados modelos de paquetes de información que brindan asistencia y revisión del diseño para especializar estos modelos en PISs y PIAs y para sometimientos específicos. Por otro lado, se desarrollan planes de migración detallada, prototipos de software y planes de pruebas para liberar implantaciones de los objetivos de migración de Administración.

Acceso. Cuenta con los servicios y funciones de soporte a los consumidores en la obtención de la existencia, descripción, localización y disponibilidad de información almacenada en el sistema de archivo, permitiendo a los consumidores solicitar y recibir documentos. Las funciones de acceso incluyen:

a) Comunicación con los consumidores para recibir solicitudes aplicando controles que limitan el acceso a la información protegida.
b) Coordinación de la ejecución de solicitudes para que se completen satisfactoriamente.
c) Generación de respuestas del estilo Paquetes de Información Diseminada (PIDs).
d) Generación de resultados y reportes para los consumidores.

4.10 Políticas y procedimientos

Dentro del contexto relacionado con el respaldo y la preservación, las políticas y los procedimientos son un conjunto de métodos que, aplicados sistemáticamente, sirven de apoyo en la realización del respaldo, el resguardo, la recuperación y la preservación de un contenido digital. Cada institución determina su propio conjunto de políticas y procedimientos aplicables sólo dentro de ella. Las siguientes son algunas políticas y procedimientos generales para el respaldo y la preservación de documentos digitales, aplicables en cualquier proyecto de digitalización.

4.10.1 Políticas y procedimientos de respaldo y preservación

1. Manejar con mucho cuidado los medios de almacenamiento.
2. Cumplir con las especificaciones del fabricante para el cuidado de los medios de almacenamiento, como las condiciones climáticas: humedad, calor, polvo, etcétera.
3. Ordenar los medios de almacenamiento en forma vertical.
4. No colocar objetos sobre los medios de almacenamiento.
5. Verificar la integridad del contenido almacenado en el dispositivo de almacenamiento secundario cada vez que se realice una copia de la información.
6. Verificar periódicamente el funcionamiento correcto del dispositivo periférico para la generación de copias de los datos.
7. Establecer reglas y procedimientos para la integración de metadatos.
8. Validar que los documentos digitales a ingresar se encuentren en un formato estándar.

4.10.2 Políticas y procedimientos de respaldo

1. Los respaldos deben hacerse en el horario de menor uso del servidor de publicación.
2. Se recomienda tener una copia del contenido digital cerca del servidor de publicación y otra lejos.
3. Retirar el medio de almacenamiento secundario de la unidad de lectura y grabación cuando haya concluido el proceso de respaldo.
4. Cumplir con los periodos de respaldo indicados en el plan de seguridad y contingencia.

4.10.3 Políticas y procedimientos de preservación

1. El acceso a la bóveda debe restringirse a un número limitado y bien definido de personas.
2. La consulta del servidor de preservación puede realizarse sólo a través del servidor de publicación de documentos digitales.
3. Refrescar los medios una vez al año.
4. Evitar el uso de los *masters* de preservación.
5. Por cada *master* de preservación generar un mínimo de dos copias.
6. Por cada *master* de publicación generar un mínimo de dos copias.
7. Asignar un límite de vida a cada documento electrónico.
8. Verificar semestralmente los cambios tecnológicos en los formatos de almacenamiento.
9. Verificar semestralmente los cambios tecnológicos de software y hardware, que impacten en la obsolescencia de los programas y los equipos en uso.
10. Verificar semestralmente la implantación de estándares nuevos en la representación y el intercambio de información digital, así como en los metadatos descriptivos y de preservación.
11. Cada vez que lo sugieran los cambios tecnológicos o los estándares nuevos, realizar la migración o el reformateo de los documentos digitales con el apoyo de programas computacionales, preferentemente.
12. Contar con un mínimo de dos bóvedas replicadas geográficamente para garantizar la preservación de los medios de almacenamiento ante desastres naturales.

5 EVALUACIÓN DE HERRAMIENTAS PARA CONSTRUCCIÓN DE REPOSITORIOS DIGITALES

En el sentido más amplio, los repositorios digitales son sitios activos y colaborativos, en los que se depositan, por diferentes personas, diversos tipos de objetos electrónicos, usualmente relacionadas por una temática o por una comunidad. Además, tecnológicamente deben facilitar la interoperabilidad para el intercambio de información con otros repositorios o con otras aplicaciones.

En la práctica, los repositorios han agrupado los contenidos de diferente forma y esto ha dado origen a una división de los repositorios (*López, Clara. 2007*), principalmente como:

- **Repositorios de eprints.** Contienen artículos científicos arbitrados o en proceso de publicación.
- **Repositorios temáticos.** Contienen objetos de un mismo tema, usualmente artículos y otro tipo de publicaciones o recursos digitales relativos al tema de interés.
- **Repositorios de materiales académicos.** No contiene sólo documentos científicos arbitrados, sino que alberga todo tipo de materiales que apoyen la enseñanza y el aprendizaje, que pueden o no corresponder al mismo tema.
- **Repositorios institucionales.** Incluyen material académico diverso, tienden a ser organizados por una institución más que por áreas temáticas. Pueden a veces incluir también documentos administrativos de la misma institución.
- **Repositorios de objetos de aprendizaje.** Basan su contenido en objetos digitales como unidades de aprendizaje, que tienen como principal objetivo transmitir un conocimiento concreto y técnicamente tienen características específicas que los hacen fácilmente reutilizables en otras aplicaciones.

5.1 Herramientas de software para construcción de repositorios digitales

En la actualidad existe un mínimo de 64 herramientas diferentes para apoyar la construcción y administración de repositorios de información digital, algunas de software libre, otras de software comercial y muchas son aplicaciones particulares.

Algunas de las características para selección de la herramienta son: Simplicidad, para instalación y uso, respaldo fuerte en soporte y documentación, esté en constante desarrollo, funcione en cualquier sistema operativo y arquitectura de cómputo, use estándares internacionales para catalogación e intercambio de información, incorpore estrategias de preservación digital y mecanismo eficiente para indización y recuperación de información, por último, acepte documentos digitales en una gran variedad de formatos.

5.1.1 Herramientas de software libre para construcción de repositorios digitales

La tabla 3 muestra un listado de diversas herramientas de software libre para construcción de repositorios digitales. Los nombres de las herramientas aparecen ordenados de forma alfabética y junto al nombre, de cada herramienta, aparece entre paréntesis el número de repositorios instalados mundialmente utilizando dicha herramienta. Ésta información se extrajo del sitio http://www.opendoar.org que aunque no es cien por ciento veraz nos puede ayudar a darnos una idea de cuántas herramientas de software libre hay para construcción de repositorios digitales y el impacto de utilización de cada una.

Puede observarse en la tabla 3 que DSpace es la herramienta de más uso con 917 instalaciones, sin embargo, en el sitio http://www.dspace.org de DSpace se observa que existen 1394 repositorios construidos. Cabe aclarar que las cifras de la tabla 3 y del sitio de DSpace corresponden al momento en que fue escrito éste trabajo.

La figura 23 muestra el porcentaje de uso de las herramientas listadas en la tabla 3. Se visualiza que DSpace ocupa el 40.7% de uso seguida por EPrints con el 14.2% (motivo por el cual se seleccionó DSpace para su evaluación).

Aunque la herramienta Greenstone aparece en la tabla 3 con 47 instalaciones es la herramienta preferida por bibliotecólogos y especialistas en tecnologías de la información por su rápida y fácil instalación. No requiere permisos especiales del sistema operativo para poder instalarse y dicha tarea la puede realizar prácticamente cualquier persona, motivo por el cuál también se seleccionó ésta herramienta para su evaluación.

Cabe aclarar que para poner en servicio un repositorio sin importar la herramienta que escojamos la instalación llega a ser bastante complicada requiriendo que ésta actividad la realice personal de cómputo altamente calificado.

No especificado (286)	Fedora (21)	panFMP (1)
Aleph (1)	Fez (9)	Perl-based (2)
Alexandria (1)	Flora (Ever-Team) (1)	Perseus (1)
ARCHIMEDES (1)	Greenstone (47)	PLEADE (1)
ArchivalWare(1)	HAL (20)	Plone (1)
arXiv (1)	HS-DVL (1)	PMB (1)
Aubrey (1)	HTML (27)	PMC (3)
Bepress (2)	Hyperwave (1)	PORT (1)
Bibdia (1)	iLisSurf (2)	Presto (1)
Caché Intersystems (1)	iLiswave-J (1)	PubMan (2)
Cadic Integrale (1)	IMIS (1)	PURE (5)
Catia (2)	InfoLib (1)	Qucosa (1)
CDSWare (2)	InfoLib-DBR (2)	RedHat (2)
Cocoon (1)	IntraLibrary (1)	refbase (1)
CONTENTdm (39)	Invenio (9)	Rhaptos (1)
Corisco (1)	IR+ (4)	SciELO(18)
CWIS (4)	Islandora (1)	Scigloo (1)
Cybertesis (13)	JaDoX (1)	SciX (3)
DARE (5)	MANITOU (1)	Shunsaku (1)
Digibib (16)	Maxwell (1)	SobekCM (2)
Digital Commons (96)	Mediatum (1)	Socionet (3)
DigiTool (19)	Mercury (1)	STAR Database (1)
Diva-Portal (32)	Micro-CDS/ISIS (2)	SWB Content (1)
dLibra (56)	Miless (1)	Symplectic (1)
DLXS (1)	MiTOS (7)	TEDE (3)
Documentum (1)	MyCoRe (9)	UBKA (1)
DOKS (5)	NALIS-R (1)	Ultraseek (1)
Drupal (5)	Nitya (1)	VITAL(16)
DSpace (917)	Nou-Rau (1)	Voorportaal (1)
Dynaweb (1)	OAICat (2)	VTS (1)
e-Repository (5)	OCS (1)	VuFind (1)
Earmas (5)	OCTOPUS (1)	WEKO (2)
eDoc (5)	OJS (3)	Wiki (1)
EPrints (321)	Open Archive (1)	WIKINDX (1)
ePub (1)	Open Repository (20)	Wildfire (1)
Equella (5)	OpenCiXbase (1)	WordPress (1)
eSciDoc (2)	OpenCourseWare (2)	XooNIps (13)
ETD-db (15)	OPUS (74)	
Exalead (1)	ORI-OAI (1)	

Tabla 3. Herramientas de software libre
(Recuperado de: [http://www.opendoar.org)](http://www.opendoar.org)

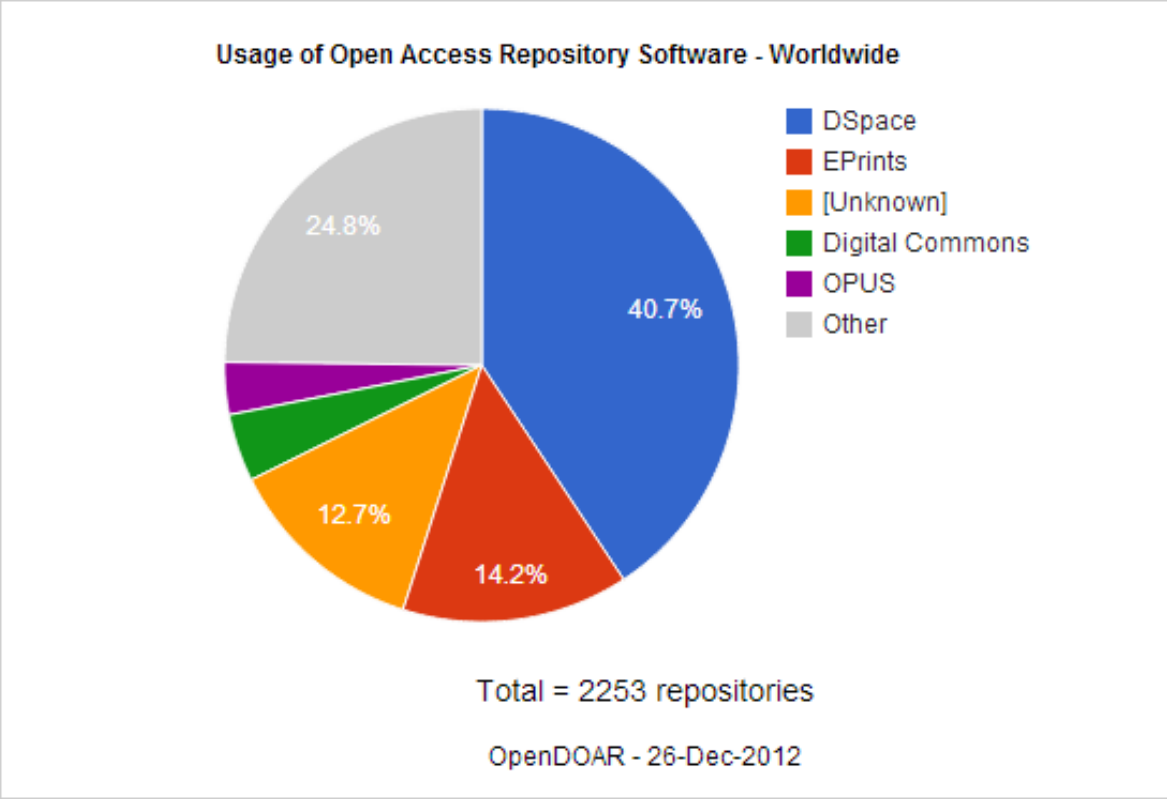


Figura 23. Herramientas más utilizadas
http://www.opendoar.org

5.1.2 Herramientas de software comercial para construcción de repositorios digitales

El número de herramientas de software comercial para construcción de repositorios digitales, comparado con las herramientas de software libre, es bastante reducido. Una característica de éste tipo de herramientas es su alto costo rebasando los 60 mil dólares ya que proveen otro tipo de funcionalidades referentes a la administración de bibliotecas. La instalación de estas herramientas también es bastante complicada pero la ventaja es que la empresa que comercializa el software es la encargada de instalarla.

Algunas herramientas de software comercial para repositorios digitales son:

1. Aleph
2. Alfresco
3. ECM Digital Asset Manager
4. Horizonte
5. Janium
6. Microsoft Sharepoint
7. Pinakes
8. Siabuc
9. Unicornio

Algunas de las herramientas, anteriormente descritas, utilizan el esquema de metadatos de MARC 21 que es un estándar más descriptivo que DublinCore puesto que posee 1000 metadatos, sin embargo, por no tener la forma de adquirir estas herramientas ni la infraestructura tecnológica para instalarlas se omitieron en la evaluación del presente trabajo, por tal motivo se seleccionaron las herramientas Alfresco y SharePoint ya que éstas sí se pueden adquirir y pueden instalarse en un equipo portátil para su evaluación.

5.2 Características de herramientas de software para construcción de repositorios digitales

En resumen la tabla 4 muestra las herramientas de software seleccionadas para su evaluación. La columna 1 describe el nombre de la herramienta. La columna 2 ilustra la versión del software al momento de realizar la evaluación, ésta es un parámetro muy importante en el proceso de evaluación ya que las versiones futuras pueden llegar a ampliar su funcionalidad ampliando con esto sus características de evaluación. La columna 3 muestra el URL del sitio que provee la herramienta y su respectiva documentación. Por último la columna 4 nos muestra la empresa o institución fabricante del software.

Software	Versión	URL	Empresa
Greenstone	2.83	http://www.greenstone.org/	Universidad de Waikato
Alfresco	Enterprise Edition 4.0.2	http://www.alfresco.com/	Alfresco Software Inc.
DSpace	1.8.2	http://www.dspace.org/	DuraSpace
SharePoint	2010	*http://sharepoint.microsoft.com/es-mx/paginas/default.aspx*	Microsoft

Tabla 4. Algunas herramientas para construcción de repositorios comerciales y libres

En el presente trabajo, los rubros que se toman en cuenta para evaluar las 4 herramientas de software de la tabla 4 son:

- Evaluación de la instalación
- Evaluación de la documentación
- Evaluación de características técnicas
- Evaluación de las funcionalidades

Para la explicación de cada uno de éstos rubros haremos referencia a una serie de tablas que a continuación son presentadas en los subtemas siguientes y que hacen referencia a cada uno de los puntos anteriores. Todas éstas tablas constan de cinco columnas que contienen como encabezado de las primeras filas el nombre de la característica que se evaluará y los siguientes cuatro encabezados contienen el nombre de la herramienta de software a evaluar.

Cabe aclarar que dichas tablas se construyen en base a pruebas hechas a cada herramienta y su objetivo es servir como una guía básica de las características más sobresalientes en la elección de dicha herramienta para la implantación de repositorios digitales. No se entrará en detalle en las pruebas de instalación y de uso, solo se abordarán los resultados en la tabla y posteriormente se describirá cada rubro de forma más detallada para cada herramienta.

5.3 Evaluación de la Instalación

Son un conjunto de características descritas en la tabla 5 que permiten conocer que tan amigable es la instalación para el usuario. Éstas características son la plataforma, el soporte y mantenimiento y las dependencias.

Plataforma.- Éste rubro se refiere al Sistema Operativo en el que puede ser capaz de ejecutarse la herramienta, es decir, nos permite conocer la estabilidad que ofrece la herramienta para ser utilizada en diferentes ambientes computacionales. Los valores que se consideraron son el(los) nombre(s) del (de los) sistema(s) operativo(s) en el(los) que funciona la herramienta o Multiplataforma, considerando para éste último valor el hecho de que la herramienta funciona perfectamente en todos los sistemas operativos.

Arquitectura.- Como su nombre lo indica, éste rubro nos permite conocer el tipo de arquitectura de cómputo para la cual está diseñada la herramienta, sus valores pueden ser SPARC, MACINTOSH, PC's y compatibles.

Soporte y Mantenimiento.- El soporte indica si se cuenta con apoyo por parte de sus desarrolladores para el proceso de instalación, administración y operación de la herramienta o si en su defecto existen comunidades que se dediquen a esto. El mantenimiento indica si la herramienta está evolucionando constantemente para corregir problemas o para ampliar sus funcionalidades. Los valores para esta característica de evaluación son: deficiente, regular, bueno, muy bueno y excelente.

Dependencias de Instalación.- Aquí se trata de explicar si la herramienta necesita de software adicional para funcionar correctamente. Ésta característica puede ser importante a la hora de elegir una herramienta respecto a las demás ya que se debe considerar las limitaciones de un equipo de cómputo para instalar herramientas adicionales y además de que pueden ser gastos extras. Lo que se busca en éste rubro es hacer un filtrado de qué herramienta es más óptima, es decir, que haga mucho más con menos. Éste rubro toma los valores de sí, en el caso de que sí necesita dependencias o no en el caso contrario.

A continuación, en la tabla 5 se presenta la evaluación de la instalación de cada herramienta:

	Greenstone	Alfresco	DSpace	SharePoint WorksPace
Evaluación de la Instalación				
Plataforma	Multiplataforma	Multiplataforma	Multiplataforma	Windows
Arquitectura	Todas	Todas	Todas	PC's y compatibles
Soporte y Mantenimiento	Muy Bueno	Excelente	Muy Bueno	Excelente
Dependencias de instalación	Sí	Sí	Sí	No

Tabla 5. Evaluación de la Instalación de algunas herramientas de software (comerciales y gratuitas) para crear repositorios digitales

Como se puede apreciar en la tabla 5 podemos ver que todas las herramientas a evaluar, a excepción de SharePoint, son multiplataforma y multiarquitectura.

Tres de ellas (Greenstone, DSpace y Alfresco) son multiplataforma por lo que funcionan en cualquier sistema operativo mientras que SharePoint únicamente se puede instalar en ambientes Windows.

Vemos también que las tres primeras herramientas son compatibles con las arquitecturas ya mencionadas anteriormente, a excepción nuevamente de SharePoint que es para PC's y compatibles.

Se observa también en la tabla que la mayoría de las organizaciones que nos proporcionan éstas herramientas les dan un buen soporte y mantenimiento, no es de esperar menos ya que éstas cuatro herramientas son de las más utilizadas para las tareas de creación de colecciones digitales por lo que deben estar garantizadas sobre cualquier tipo de falla, sin embargo las dos herramientas de tipo comercial (Alfresco y SharePoint) tienen la ventaja de que existe por parte de la misma organización un soporte comercial que brinda muchas más opciones de contacto entre el usuario y los técnicos específicos expertos en el área para la solución de prácticamente cualquier problema, es por ésta razón que a ambas herramientas comerciales, se les evaluó con excelente, en comparación a las dos herramientas libres.

Las tres primeras herramientas descritas en la tabla tienen dependencias hacia otras herramientas de software para su completa implantación y además les brindan mayor usabilidad.

Para la instalación de Greenstone se necesita (La instalación y configuración de cada dependencia de software es automática y muy sencilla para modo de

pruebas y evaluación, sin embargo pensar en montar un repositorio digital o una colección de éstos es sumamente complejo y demanda altos conocimientos en TI):

- Servidor Web Apache
- Perl
- GCC
- GDBM
- Compilador Java (versión 1.4.0 o superior)

Por otro lado, Alfresco necesita obligatoriamente (todo esto lo instala la herramienta de forma automática):

- Java SE Development Kit (JDK)
- Servidor de Aplicaciones (Tomcat, JBoss o WebLogic)
- Base de Datos (PostgreSQL, Oracle, SQL Server, MySQL o DB2)

Instalaciones Opcionales para Alfresco y Greenstone (se instalan manualmente):

- OpenOffice
- ImageMagick
- Flash Player (Version 10.x)
- SWF Tools (pdf2swf)

Para la instalación de DSpace se necesita (la instalación y configuración demanda un alto conocimiento en computación y en TI):

- Oracle Java JDK 6 (El SDK estándar está bien, no se necesita el J2EE, Java 7 actualmente no es soportado).
- Apache Maven 2.2.x o superior (Java build tool).
- Apache Ant 1.8 o posterior (Java build tool).
- Base de Datos Relacional (PostgreSQL 8.3 a 8.4 u Oracle 10g o superior).
- Motor de Servlets: (Apache Tomcat 5.5 o 6, Jetty, Caucho Resin o equivalente). - Oracle Java JDK 6 (El SDK estándar está bien, no se necesita el J2EE, Java 7 actualmente no es soportado).

Por último, al hablar de las dependencias de la herramienta SharePoint de Microsoft se observa en la tabla 5 que no requiere ninguna dependencia de software para su instalación (se instala con el asistente de la paquetería de Office), sin embargo, vale la pena mencionar que SharePoint permite ampliar sus funcionalidades si se instala el SharePoint Server, para éste último se necesita:

- SQL Server back-end para Groove Server Manager (Microsoft SQL Server 2008 R2, SQL Server 2008 con Service Pack 1 (SP1) y Cumulative Update 2 o SQL Server 2005 con SP3 y Cumulative Update 3).

Ésta última herramienta, únicamente tiene dependencia hacia software de Microsoft. Aunque SharePoint WorksPace 2010 no tiene dependencias de software se puede combinar con SharePoint Server 2010 (SP SERVER) para extender sus posibilidades de uso, sin embargo, SP SERVER tiene como dependencias para trabajar de forma correcta la herramienta descrita anteriormente y que además es propiedad de Microsoft, como ya se había mencionado antes.

Con todo lo anterior se aprecia también que la herramienta Alfresco, de distribución comercial, comparte características (de dependencia a otros software) con las herramientas de libre distribución ya que necesita, por ejemplo, un servidor de aplicaciones (Tomcat, JBoss o WebLogic) de distribución libre, así como otras herramientas que también lo son. Así mismo se aprecia que la herramienta Alfresco requiere de una serie de instalaciones opcionales para extender sus posibilidades.

5.4 Evaluación de la Documentación

Son un conjunto de características que permiten conocer la facilidad que tiene el usuario para documentarse sobre la herramienta. Se califica el idioma y la buena traducción entre éstos, la sencillez de los propietarios para dar a entender las opciones de su producto y demás características que ayuden a que el usuario sea capaz de operar la herramienta sin necesidad de recurrir a situaciones que puedan generar gastos extras. Éstas características son la documentación para la instalación, para el administrador y para el usuario.

Documentación para instalación.- Es la característica que permite conocer que tan buena es la documentación que la empresa propietaria de la herramienta a evaluar proporciona a sus usuarios para poderla instalar de una forma sencilla y directa. Se evalúa con deficiente, regular, bueno, muy bueno y excelente.

Documentación para administrador.- Característica que permite conocer que tan buena es la documentación que la empresa propietaria de la herramienta a evaluar proporciona a sus usuarios para poderla operar al grado de administrador. Se evalúa con deficiente, regular, bueno, muy bueno y excelente.

Documentación para usuario.- Característica que permite conocer que tan buena es la documentación que la empresa propietaria de la herramienta a evaluar proporciona a sus usuarios para poderla operar al grado de usuario. Se evalúa con deficiente, regular, bueno, muy bueno y excelente.

A continuación, en la tabla 6 se presenta la evaluación de la documentación de cada herramienta:

	Greenstone	Alfresco	DSpace	SharePoint WorksPace
Evaluación de la Documentación				
Documentación para instalación	Buena	Muy Buena	Buena	Muy Buena
Documentación para administrador	Buena	Muy Buena	Buena	Muy Buena
Documentación para usuario	Buena	Muy Buena	Buena	Muy Buena

Tabla 6. Evaluación de la Documentación de algunas herramientas de software (comerciales y gratuitas) para crear repositorios digitales

Algo que comparten todas las herramientas es que al menos coinciden en que su documentación debe brindarse en lenguaje inglés, la única herramienta que tiene soporte para más lenguajes es Greenstone ya que tiene soporte en español, francés, ruso, portugués, inglés, árabe, vietnamita y kazajistán.

También, y en mi opinión como es de esperarse, las herramientas de soporte comercial tienen una muy buena documentación (en comparación a las herramientas de distribución libres) que cubre los rubros que se han evaluado, cuentan además (desde las páginas oficiales de sus organizaciones) con una amplia gama de videos relacionados a éste tipo de tareas.

Mucha de la documentación de las herramientas de software libre (en relación al soporte que se les da a éste tipo de herramientas) se encuentra en foros, en wikis y en ocasiones videos puestos por practicantes de las TI en la web y aunque a veces toda esta información está en inglés es bastante buena para poder manipular la herramienta libre a un buen nivel.

5.5 Evaluación de las Características Técnicas

Éstas características describen el aspecto técnico de la herramienta. Se describen: Las estrategias de preservación, el almacenamiento de información, los protocolos intercambio de información, los esquemas de metadatos, los lenguajes de representación de contenidos, el indizador, los tipos de objetos, el lenguaje de prog. en el que está hecho y sus módulos de aplicación.

Estrategias de Preservación Digital.- Lista las técnicas de preservación, descritas en el *capítulo 4 y que* la herramienta utiliza.

Almacenamiento de Información.- Describe qué tecnología se utiliza para almacenar información: Sistema de admon. de archivos o sistema de admon. de bases de datos.

Protocolos de intercambio de información.- Lista los protocolos de información (*véase capítulo 3 para más información*) que la herramienta utiliza. Estos protocolos de intercambio de información son fundamentales a la hora de evaluar el que una herramienta permita el importado y exportado de información y que tan bien lo hará ya que es por medio de estos protocolos el que una herramienta realiza las dos actividades ya mencionadas y que son muy esenciales para los repositorios de hoy en día.

Esquemas de metadatos.- Lista el esquema de metadatos (*véase capítulo 2 para más información*) que la herramienta utiliza.

Lenguajes de representación de contenidos.- Lenguajes que se utilizan para representar internamente la información.

Indizador.- Lista el(los) nombre(s) del(de los) software(s) de indización que usa la herramienta.

Formatos.- Hace referencia al tipo de formatos de los objetos digitales que se pueden archivar en el repositorio.

Lenguaje de prog. en el que está hecho.- Especifica el nombre del lenguaje en el que está hecha la herramienta. Ésta característica nos puede ayudar a conocer el nivel de portabilidad de la herramienta. Si está hecho en JAVA, nos indica que la herramienta es multiplataforma.

Módulos de aplicación.- Lista herramientas de software adicionales para mejorar sus funcionalidades como plugins, etc.

A continuación, en la tabla 7 se presenta la evaluación de las características técnicas de cada herramienta:

	Greenstone	Alfresco	DSpace	SharePoint WorksPace
Evaluación de las Características Técnicas				
Estrategias de Preservación Digital	Rejuvenecimiento, Emulación, Estandarización, Migración, Encapsulado, Autenticidad	Reformateo, Replicación, Estandarización, Migración, Encapsulado	Autenticidad, Encapsulado, Rejuvenecimiento, Estandarización	Migración, Actualización, Replicación.
Almacenamiento de Información	GDBM	JMX Oracle, Postgres	Oracle, Postgres	Administrador de archivos de Windows
Protocolos de intercambio de información	OAI-PMH, Z39.50, SRW	HTTP(S), SOAP.	OAI-PMH, Herramienta para Z39.50	Data lock-in, IPv6 o IPv4 o ambos.
Esquemas de metadatos	Dublín Core	CMIS y Dublín Core	Dublín Core	Dublín Core
Lenguajes de representación de contenidos	XML, METS	XML, HTML.	METS, XML	Visual Studio 2010, XML, JavaScript, ASP.NET
Indizador	MCPP, MG, Lucene	Lucene	Lucene	Windows Search
Extensión de archivos	pdf, postscript, rtf, html, txt, latex, excel, ppt, email, código fuente, open office, gif, jif, jpeg, tiff, mp3, ogg vorbis, midi, zip	css, xml, html, rtf, java class, java script, audio de cualquier tipo, 3g, 3g2, adobe (acrobat xml, aftereffects, air, flex, framemarker, ilustrator, indesign, pagemarker, pdf, photoshop, premiere, soundbooth), microsoft office (ms-word, ms-excel, ms-powerpoint, ms-project, ms-outlook),iwork keynote, iwork numbers, iwork pages, autocad openoffice, email	css, xml, html, rtf, jpeg, gif, bmp, png, tiff, photo-cd, mpeg, x-aiff, x-wav, x-pn-realaudio, mpeg, quicktime, marc (machine readable cataloging records), mathematica, microsoft office (ms-word, ms-excel, ms-powerpoint, ms-project, ms-visio), pdf, postscript, sgml, wordperfect, tex-dvi, fmp3, latex, photoshop, tex-document	office, mapa de bits, flash, .library-ms, .rar, css, xml, html, rtf. jpeg, gif, bmp, png, tiff, photo-cd, mpeg, x-aiff, x-wav, x-pn-realaudio, mpeg, quicktime, pdf
Lenguaje de prog. en el que está hecho	C++, Java	Java	Java	C++, C#.
Módulos de aplicación	Plugins	Plugins	Plugins	Plugins

Tabla 7. Evaluación de las Características Técnicas de algunas herramientas de software (comerciales y gratuitas) para crear repositorios digitales

Hablando sobre las estrategias de preservación digital que ofrecen las cuatro herramientas Greenstone nos ofrece las seis que son mencionadas en la tabla 7, haciendo mención de que la migración nos la permite hacer hacia DSpace.

Alfresco solo nos ofrece las cinco estrategias de preservación digital que se exponen anteriormente en la tabla.

Por otro lado DSpace, aparte de incluir las técnicas de la tabla anterior, también nos permite el manejo de versiones y flexibilidad para agregar metadatos de preservación.

Y finalmente SharePoint nos ofrece como se ve en la tabla solamente tres técnicas de preservación digital, con lo que se concluye que las herramientas de distribución libre dan más posibilidades al usuario en éste rubro.

El almacenamiento de la información de las herramientas DSpace y Alfresco es a través de bases de datos como por ejemplo Oracle y Postgres, para el caso de SharePoint Server se utiliza también un DBMS que es SQL Server R2, sin embargo para el caso de SharePoint Workspace y Greenstone se utilizan sistemas de administración de archivos.

Sobre los protocolos de intercambio de información (*véase capítulo 3 para más información*) se puede ver que todas las herramientas utilizan ya algunos de los protocolos más modernos como lo son HTTPS y los IPv6 e IPv4 por ejemplo.

Vale la pena hacer mención en éste rubro que el protocolo OAI-PMH le permite a ambas herramientas de software libre importar y exportar información.

También, al hablar de SharePoint Workspace se toma en cuenta que para los protocolos de intercambio de información, utiliza Data lock-in (a excepción de entradas individuales cuando utiliza SSTP usando TCP y HTTP) y cuando se implementa SharePoint Server se usan IPv6 o IPv4 o ambos.

Dentro del rubro que hace referencia al esquema de metadatos se aprecia de la tabla que el esquema de metadatos que más se utiliza es el de Dublín Core (Metadatos Descriptivos), Si se hace referencia al capítulo dos se comprueba que éste esquema de metadatos es de los más usados actualmente para manipular objetos de un repositorio digital.

Los lenguajes de representación de contenidos son muy parecidos a excepción de SharePoint ya que utiliza lenguajes de su misma organización (Microsoft). También no es de sorprender que un lenguaje ya muy universal e indispensable como lo es XML se encuentre disponible para las cuatro herramientas evaluadas.

El indizador de estas herramientas es una característica muy importante y de vital importancia en la hora de la elección de una herramienta ya que la velocidad de tiempo que tarda en indizar es muy variada y muy importante. Las herramientas (Alfresco y DSPace) utilizan un indizador Lucene y da muy buenos resultados. Para el caso de Greenstone el indizado se hace mediante MCPP siendo muy deficiente pues los tiempos de velocidad de indizado tardan demasiado.

Por otro lado, la herramienta comercial SharePoint Workspace aprovecha la capacidad de indizar del propio ambiente de Windows sobre el que está instalado (Windos Search para Windows 7) y mediante IFilters, que es la manera en la que toda la paquetería de Microsoft Office indiza. Vale la pena mencionar también que Share Point Server indiza exactamente de la misma manera que Workspace, sin embargo se le puede ampliar ésta función con la instalación de Sharepoint Fast Search Server utilizando el motor de búsquedas FAST quien permite realizar búsquedas mucho más complejas sobre un indizado más completo.

Estas cuatro herramientas cumplen muy bien su característica de que son herramientas para crear colecciones digitales pues permiten almacenar casi cualquier tipo de archivo digital como fotos, videos, audios, documentos y demás, lo interesante es ver también las extensiones de los tipos de archivos en los cuales están limitadas (*véase la tabla 7*). Esto también va de la mano con la capacidad que tienen las herramientas evaluadas de poder implantar y/o conectarse con otros tipos de software que les sean capaces de ampliar sus funcionalidades como por ejemplo por medio de módulos de aplicación como son los plugins y donde se ve en la tabla que todas las herramientas los ofrecen, por ejemplo Greenstone requiere de plugins para pdf, postscript, word, exel, ppt, zip jpeg, mp3, html, latex, Alfresco por su parte necesitaría plugins para pdf, html, ajax, flash, kofax (digitalización de producción). DSpace por ejemplo, necesitaría Plugins para indizado y por último SharePoint puede usar Plugins (ej. para mejor búsquedas, integración con java, etc.), en fin como ya se mencionó todas las herramientas ofrecen una amplia gama de formatos para el almacenamiento de sus objetos.

Los lenguajes de programación con los que están hechas las herramientas no difieren mucho, salvo por la excepción de SharePoint que está programado en C++ y C# (como las herramientas de la familia de Office que ofrece Microsoft

Windows) y Greenstone que está programado en su mayoría en C++ y con una GLI desarrollada en java, para las demás herramientas el lenguaje utilizado es Java (de aquí su clara dependencia a la instalación de éste software para su funcionamiento).

Ésta característica puede diferenciar una herramienta de otra pues para extender sus posibles funciones conocer el lenguaje de programación en el que se desarrolló es de vital importancia ya que de aquí se parte en la investigación de la facilidad de integrar la herramienta con otros lenguajes para darnos libertad de desarrollo, además se debe contemplar al momento de seleccionar un lenguaje para desarrollar una herramienta para gestión de repositorios digitales conceptos simples como: utilizar el más moderno, el más barato, el más usado, el más sencillo, el más estable… etc.

5.6 Evaluación de las Funcionalidades

En este apartado se evalúan las características que la herramienta ofrece para la interacción con el usuario. Se pueden apreciar como rubros de éste apartado los siguientes: Eficiencia, Tipo de Búsquedas, Importación de la Información, Exportación de la Información, Estadísticas de uso, Manejo de múltiples colecciones, Personalización de la Interfaz, Manejo de, Perfiles de Usuario, Tipo de recurso electrónico, Colecciones distribuidas y las Novedades de cada herramienta.

Eficiencia.- Se refiere al hecho de que el sistema use los recursos de cómputo adecuadamente y que los tiempos de respuesta sean razonables. Sus valores son Sí o No.

Tipo de Búsquedas.- Describe la forma en la que la herramienta permite realizar búsquedas de información lo que permite filtrar por facilidad y rapidez de búsqueda. Solamente toma los valores de Tipo de Búsquedas por Campo de Texto (desde un campo de texto se especifica el término o frase a buscar) y/o por Clasificador (clasificaciones por listas alfabéticas: de autor, de tema, de título, de palabra clave, etc)

Importación de la Información / Exportación de la Información.- Hace referencia a que si la herramienta permite importar / exportar datos. Sus valores son Sí o No.

Estadísticas de uso.- Característica que describe si la herramienta realiza de forma natural estadísticas sobre sus datos. Sus valores son Sí o No.

Manejo de múltiples colecciones.- Describe si es posible que la herramienta maneje más de una colección. Sus valores son Sí o No.

Personalización de la Interfaz.- Describe si es posible que la herramienta pueda modificar su interfaz. Sus valores son Sí o No.

Manejo de Perfiles de Usuario.- Describe si es posible que la herramienta maneje más de un perfil de usuario. Sus valores son Sí o No.

Formato.- Indica el tipo de formatos que la herramienta puede manejar. Los posibles valores de este rubro son Texto, Imágenes, Audio, Video y/o Aplicaciones.

Catálogo de autoridades.- Nos describe si la herramienta ofrece un vocabulario controlado (o catálogo) en la asignación de metadatos para evitar que los usuarios cometan errores de captura en las etiquetas descriptivas de los objetos de sus colecciones mejorando con esto el formato de la entrada al repositorio de estos objetos. Sus valores son Sí o No.

Colecciones distribuidas.- Se refiere a si la herramienta es capaz de sostener múltiples colecciones distribuidas en espacios físicamente distintos. Sus valores son Sí o No.

Novedades.- Describe características novedosas de la herramienta.

A continuación, en la tabla 8 se presenta la evaluación de las funcionalidades de cada herramienta:

	Greenstone	Alfresco	DSpace	SharePoint WorksPace
Evaluación de las Funcionalidades				
Eficiencia	No	Sí	Sí	Sí
Tipo de Búsquedas	Campo de Texto y Clasificador	Campo de Texto y Clasificador	Campo de Texto y Clasificador	Campo de Texto y Clasificador
Importación de la Información	Sí	Sí	Sí	Sí
Exportación de la Información	Sí	Sí	Sí	Sí
Estadísticas de uso	No	No	Sí	No
Manejo de múltiples colecciones	Si	Sí	Sí	Sí
Personalización de la Interfaz	Sí	Sí	No	No
Manejo de Perfiles de Usuario	Sí	Sí	Sí	Sí
Tipo de recurso electrónico	Texto, Imágenes, Audio, Video y Aplicaciones.	Texto, Imágenes, Audio, Video y Aplicaciones.	Texto, Imágenes, Audio, Video y Aplicaciones.	Texto, Imágenes, Audio, Video y Aplicaciones.
Catálogo de Autoridades	No	No	Sí	No
Colecciones distribuidas	Sí	Sí	Sí	Sí
Novedades	-	Disponible para móviles. Acceso in the cloud.	Permite habilitar el protocolo HTTPS.	Disponible para móviles. Acceso in the cloud.

Tabla 8. Evaluación de las Funcionalidades de algunas herramientas de software (comerciales y gratuitas) para crear repositorios digitales

Para evaluar un poco la calidad de las cuatro herramientas seleccionadas, de acuerdo a la definición anterior de eficiencia se afirmó que solo Greenstone no es eficiente, esto se ve reflejado principalmente en su técnica de indizado ya que demora mucho y para la utilización de un repositorio serio y totalmente funcional este modo de indización no es de utilidad. Evaluando bajo este mismo rubro a las otras tres herramientas, sí son en alto grado muy eficientes.

Gracias a los potentes indizadores comentados en las líneas anteriores que ofrecen todas las herramientas de software que se han descrito es posible realizar búsquedas muy complejas. En la tabla 8 se aprecia que todas las herramientas ofrecen tipos de búsquedas por campo de texto (complejas, pues permiten incorporar comodines y operadores lógicos aplicados sobre expresiones lógicas que combinen diversos metadatos como autor, título, palabras claves, rangos de fechas en el caso de SharePoint) y también búsquedas por clasificador, es decir, por listas alfabéticas: de autor, de tema, de título, de palabra clave, etc.).

Otras dos características muy importantes en la elección de herramientas pueden ser el hecho de que éstas permitan la importación y exportación de información e incluso que puedan importar y exportar hacia otras herramientas también conocidas (como lo son todas éstas). Tal es el caso de las herramientas SharePoint (mediante interfaz de Windows, powershell en formato .cmp) y Alfresco (En AIPs que usan METS XML para metadatos) que ofrecen exportación de su información entre ambas herramientas.

Una función muy atractiva es la característica de poder generar reportes de estadísticas de uso sobre la información que involucra a una colección o repositorio digital, siendo estrictos solo Dspace ofrece esta posibilidad de una forma muy limitada, generando bitácoras de uso del repositorio y nos ayuda a muestrear el número de veces que ha sido descargado un objeto del repositorio e incluso nos da la estadística a nivel países, estados y hace un corte por años y meses, por otro lado SharePoint Workspace permite, mediante implementación con SharePoint Server, hacer:

- Informes de administración (tasas de rastreo por origen de contenido, tiempo de los resultados de las búsquedas)
- Informes de Web Analytics

Para el caso de Alfresco, ninguna de éstas características está disponible al inicio, pero resulta factible implantar todas éstas funciones, lo que ocurriría quizá también con DSPace y con SP SERVER para ampliar la cantidad de informes estadísticos generados.

Es claro que todas las herramientas permiten manejar múltiples colecciones, sobre la personalización de sus interfaces, Greenstone y Alfresco sí permiten como una de sus tareas comunes la modificación de la interfaz (aunque es muy burda la personalización que ofrecen) con muy poca configuración por parte del usuario, además Alfresco (que trabaja en web) permite una modificación más al gusto del usuario modificando el código de programación. DSpace de forma inicial nos ofrece una interfaz por omisión que puede ser modificada cambiando las propiedades del sistema y modificando el código, el caso de SharePoint Workspace es similar al de DSpace ya que no es posible personalizar más allá de lo que la familia de Office nos permite hacer, sin embargo. SP SERVER (que trabaja en web) sí es de interfaz modificable alterando el código. Con todo esto, se rescata también que las herramientas que trabajan en web son de interfaz modificable alterando su código de programación.

En cuanto a los perfiles de usuario todas las herramientas permiten manejar más de un perfil, aunque solo nos encontramos con los clásicos administrador y usuario para el caso de las cuatro herramientas. Greenstone, por ejemplo, para el acceso a las colecciones digitales maneja dos tipos de usuario que son Administrador y usuario, sin embargo, para la administración del repositorio maneja Usuario y Bibliotecario (Administrador). Para el acceso a los documentos Greenstone acepta dos tipos de permisos: Permiso de acceso a toda la colección (con cuenta y contraseña) y permiso de acceso a un documento (con cuenta y contraseña), esto es que se nos permite acceso con protección con cuenta y contraseña a ciertas colecciones en los modos anteriores.
DSpace por otro lado, permite por default dos tipos de perfil, el de administrador y usuario, pero también permite el manejo de cualquier tipo de usuarios creados desde la interfaz (grupos y tipos de usuarios). Alfreso y SharePoint funcionan casi de manera similar a DSpase en la cuestión de que nos permiten crear grupos de usuarios para las distintas colecciones y asignar a las colecciones de los repositorios los perfiles de público o privado para el acceso a dicha información.

El tipo de los recursos electrónicos son en general los que se describieron en la tabla 8.

Para el rubro del catálogo de autoridades (vocabulario controlado) se aprecia que solamente la herramienta DSpace es la única que ofrece ésta característica lo que ayuda enormemente a tener un mejor control y clara organización sobre los objetos que entran al repositorio en comparación a las otras tres herramientas. Como ejemplo se menciona el catálogo de autoridades de DSpace para el lenguaje, lo que permite elegir al usuario uno de entre varios para la interacción con los objetos del repositorio.

Las colecciones distribuidas se intercambian mediante protocolos como Corba para Greenstone y NetBEUI para SP Workspace y URL para las demás herramientas incluyendo SP Server.

Algunas de las novedades para estas herramientas (excluyendo a Greenstone) son que ya están incluso disponibles para dispositivos móviles (cosa que es muy novedosa en estos tiempos) y que también ya cuentan con acceso a la nube como Alfresco y SharePoint, otras como DSPace ofrecen la posibilidad de habilitar el protocolo HTTPS.

También podemos afirmar de manera contundente que todas estas herramientas son las de mayor difusión y uso, todo esto reflejado en que han sido elegidas por instituciones del más alto prestigio y por su fácil alcance monetario (para el caso de las herramientas comerciales).

6 CONCLUSIONES

El uso de una herramienta como lo es un repositorio digital que nos permite realizar actividades como administración, organización, preservación y acceso de información, etc. es cada día más esencial. A lo largo de éste trabajo se describieron las bases que tecnológicamente se requieren para implantar éste tipo de herramientas.

Ahora se sabe que el manejo de archivos es una característica muy importante así como el conocer un poco sobre la teoría de las bases de datos pues nos permite tener una idea de donde almacena información una colección digital en línea. Así mismo el conocer sobre los índices nos ayuda a entender mejor el proceso que tiene el recuperado de la información a la hora de hacer consultas de búsquedas desde la interfaz de usuario.

También, el concepto de los metadatos (datos sobre datos) es uno de los que más presencia tiene ya que su uso nos permite (basándonos en estándares) catalogar de alguna manera la información que va a entrar a nuestro repositorio, aquí se ve claramente que los conceptos van ligados pues la buena asignación de metadatos van de la mano con la buena y óptima búsqueda y recuperación de información.

Si se habla de los protocolos que utiliza un repositorio para intercambiar información se está haciendo referencia a la manera en la que es capaz de importar y/o exportar datos, a la manera en la que permite comunicar comunidades de datos digitales, la forma en la que permite conectar varios computadores de una red al proceso de trabajo colectivo en línea. Por citar algunos existen los protocolos HTTP (que permite transmitir información basado en web), Z39.50 (que permite buscar y recuperar información en bases de datos), SOAP (que permite intercambiar información a través de HTTP y por medio de XML) y el SRW (que permite hacer consultas en internet). Aunado a esto se relacionan los conceptos de redes y topologías de datos., así como arquitecturas de redes y modelos de referencia.

Técnicas como la migración, el reformateo, la replicación y la estandarización son de mucha ayuda en el tema de preservar los documentos digitales alojados en el repositorio.

Abordando la evaluación de las herramientas de construcción de repositorios tanto públicas como privadas me es muy grato concluir que todas cumplen las expectativas que las hacen ser de las más solicitadas y utilizadas por numerosas y grandes instituciones.

Para la elección de la mejor herramienta vería más que nada aquella que hiciera más y a bajo precio. Baso mi elección en cuestiones tales como: tipo de ambientes que yo sé usar, la respuesta en cuanto a menor tiempo de búsqueda y recuperación que ofrecen las herramientas, el soporte que se le da a la

herramienta, el idioma y todo esto en conjunto con las descripciones que se abordaron en las tablas de evaluación anteriores.

Yo considero que la herramienta que realmente cubre con todos estos rubros es DSpace ya que es una herramienta de libre licencia lo que hace que se ahorre un gasto por mínimo que sea, como lo es con Alfresco que tan solo cobra 10 dólares mensuales por usuario para conectarse a la nube con 10 GB de almacenamiento y con soporte comercial, Dspace es además muy utilizada por grandes instituciones ya que tiene en su repertorio mil trescientas cincuenta y un instituciones serias, entre ellas la Facultad de Filosofía y Letras de la UNAM.

Antes de hablar de la evaluación de la instalación de las herramientas para creación de repositorios digitales, se debe de tomar mucho en cuenta la infraestructura que se tiene, es decir, se debe considerar antes de instalar una herramienta de este tipo aspectos como el personal humano para operarlos (administradores, programadores, etc) quienes se deberán encargar de revisar los prerrequisitos que demanda la instalación de la herramienta y ejecutarlos, encargados para actualizar las herramientas y operarlas, mantener los equipos de cómputo en óptimo estado en sus variantes tanto de software como de hardware (revisando en ocasiones la posible migración de datos a equipos de cómputos más modernos), en fin, se sugiere que se tome en cuenta que el hecho de instalar un repositorio digital conlleva a una serie de revisiones y estudios (desde la puesta en marcha de un repositorio hasta su operación) sobre recursos humanos (diferentes roles de personas involucradas) y tecnológicos (computadores, servidores, redes, etc.) y que además también cuentan como prerrequisitos que deberá tomar en consideración la institución que requiera el uso de un repositorio digital.

Dentro de la evaluación de la instalación y conforme a lo que se describió en la tabla 5, es muy claro que las primeras tres herramientas: Greenstone, Alfresco y DSpace tienen la mejor calificación pues brindan la mayor portabilidad de uso pues son multiplataforma, prácticamente se instalan en todas las arquitecturas y cuentan con un muy buen soporte y mantenimiento, sin embargo, es la herramienta DSpace la que mejor mantenimiento tiene pues aun siendo una herramienta de licencia libre se le da mucho seguimiento por parte de innumerables comunidades, además es debido a su filosofía de libre distribución que sus actualizaciones avanzan muy rápido, lo que considero una gran ventaja contra una herramienta de distribución comercial. Por otro lado cuando hablamos de las dependencias de la instalación para las herramientas, es claro que DSpace necesita muchas herramientas que demandan un alto conocimiento en TI, sin embargo, se hace la distinción con todas las herramientas que se evaluaron en el tipo de instalación que se pueden realizar llegando a dos distinciones, es decir, se tienen la instalación de prueba y la instalación para el servicio, en donde Greenstone, Alfresco y SharePoint difieren en grado de dificultad para cada tipo de instalación, a simple vista de la tabla 5 y de su posterior descripción se ve que son herramientas de muy sencilla instalación para su prueba, pero si se quisieran instalar estas herramientas para un uso más serio, más formal, resultan más

complejas sus instalaciones, a diferencia de esto, DSpace mantiene su grado de dificultad.

Es de considerar también que a las herramientas de licencia comercial se les da un soporte excelente y si se quisiera trabajar con alguna de ellas y ponerlas en servicio es evidente que el usuario tendría mínima o nula participación en dicha instalación, todo lo haría el proveedor. Esta es una clara ventaja que tienen las herramientas de software comercial ante las de licencia libre, sin embargo habría que considerarse también y como ya se mencionó que los costos de estas herramientas son muy elevados.

DSpace también nos ofrece un muy buen soporte poniendo a disposición de todos los usuarios wikis, foros y formularios de contacto así como documentación oficial (en inglés) del uso y administración de la herramienta. Una desventaja puede ser el hecho de que están disponibles la mayoría de sus utilidades (manuales, portal e incluso la herramienta) en el lenguaje inglés.

Otra desventaja menor que tiene a comparación de las otras herramientas en cuanto al soporte es que la mayoría de las cosas se programa en java y lenguajes orientados a la web, por lo que si un usuario no tiene experiencia en estas tecnologías, puede llevar a un gasto extra al contratar un experto en TI para la solución o implantación de extensiones de la misma herramienta. Sin embargo esta situación es muy similar en las otras herramientas. Incluso partiendo desde ese caso, yo considero que (al menos para los usuarios de las TI) una gran ventaja es el uso de esos lenguajes de programación ya que son los que están predominando en la actualidad lo que podría facilitar el desarrollo de extensiones para la herramienta.

Cuando se evalúa la documentación, mediante la tabla 6, se indica que las herramientas de licencia comercial están por encima de las herramientas de licencia libre, esto por el hecho de que las empresas proveedoras de estas herramientas tienen todo un equipo de personas expertas en TI dedicadas exclusivamente a la redacción de las documentaciones que sirven de apoyo a los usuarios finales para el uso del software. Tanto Greenstone, como DSpace tienen mucha documentación en diferentes idiomas y ofrecen muchas alternativas para dar solución a problemas que puedan llegar a tener los usuarios, sin embargo, la experiencia en la instalación de DSpace haciendo uso del manual que ofrecen para esta etapa no fue muy buena, la documentación que existe para el uso de las herramientas tanto como administrador y usuario de las dos herramientas comerciales son excelentes y no hubo mucho problema, para Greenstone y DSpace se indagó en muchos sitios web y algunas de las otras opciones que se ofrecen para poder operar estas herramientas como usuarios y aunque realmente no son tan deficientes estas ayudas son una desventaja en relación al otro tipo de soporte que dan las herramientas de licencia pagada, lo que reitera lo mencionado anteriormente del buen soporte que se da en este tipo de herramientas en comparación a las de libre licencia.

En la evaluación de las características técnicas con ayuda de la tabla 7 vemos que Greenstone, Alfresco y nuevamente DSpace proporcionan al usuario más estrategias de preservación digital. También vale la pena comentar el hecho de que la mayoría de las herramientas de distribución comercial y de mayor valor monetario no proporcionan este tipo de estrategias por lo que es una gran desventaja en comparación con las herramientas de software libre. También, ninguna herramienta de las evaluadas se apega al modelo de referencia OAIS, sin embargo es el software DSpace quien más lo hace.

El almacenamiento de la información como ya se comentó antes, es un factor muy importante y que relaciona temas igual de importantes como lo son la recuperación de la información. En este apartado las dos mejores herramientas son Alfresco y DSpace pues almacenan la información haciendo uso de los DBMS ya descritos y obviamente su recuperación de información es más completa.

De los Protocolos de intercambio de información es rescatable mencionar que es por medio de estos que las herramientas se distinguen en que tan buenas son para la importación y exportación de datos. De la tabla 7 vemos que tanto Greenstone como DSpace utilizan OAI, protocolo que está siendo de mayor uso en la actualidad.

Otro paréntesis que se hace es en el apartado que evalúa los esquemas de metadatos ya que el mejor esquema de metadatos dada su alta representatividad es MARC21, sin embargo ninguna de las herramientas evaluadas maneja este esquema, sin embargo vale la pena también mencionar que el uso de este esquema de metadatos es una característica de las herramientas de software comercial, lo que a su vez representa una gran desventaja sobre las herramientas de software libre. Por otro lado todas las herramientas que se evaluaron manejan el esquema Dublín Core cuyo uso es cada día más frecuente y más universal.

En cuanto a los lenguajes de representación de contenidos, la tabla 7 nos dice que uno muy común es el lenguaje XML, esto es porque su uso es cada día más esencial y de nuevo DSpace sobresale en este apartado.

Como se ha venido mencionando antes, el Indizador es un factor vital para la puesta en marcha de un repositorio, evaluando la velocidad del indizado se concluye que es DSpace quien mejor realiza la tarea en comparación a Greenstone, Alfresco y también SharePoint.

De igual manera sucede en los formatos que soportan las herramientas, se ve claramente el hecho de que son DSpace y Alfresco quienes más formatos admiten y nuevamente se ve el potencial que tiene la herramienta de software libre de estar a nivel de las herramientas comerciales.

La portabilidad que tienen Dspace y Alfresco está ligada al lenguaje de programación en el que está hecho (desarrollados en lenguaje java). Esto considero que es también algo muy útil ya que estamos hablando de que son

herramientas multiplataforma, sin embargo, una plataforma e interfaz más sencilla y a la que la mayoría (creo yo) de los usuarios está acostumbrado es la que ofrece Windows en office, por lo que SharePoint en éste rubro también tiene una mención especial. También es muy útil para una herramienta el estar diseñada (para su fácil readaptación con códigos de otros desarrolladores) bajo un lenguaje de programación eficiente, sencillo y de alto uso y más si planea brindar la opción de poder extender sus funcionalidades y aunque la mayoría de las herramientas de software comercial no nos lo permiten las dos que evaluamos sí lo hacen (solo cuando hablamos de SharePoint Server) y los lenguajes de programación en los que están hechas las herramientas evaluadas cumplen con ser sencillos y de un muy alto uso por muchos usuarios.

Para terminar la conclusión de esta evaluación de características técnicas se observa también de la tabla 7 que las cuatro herramientas ofrecen la posibilidad de ampliarse mediante plugins lo que resulta muy bueno si se desea realizar tareas más demandantes dentro de los objetos de un repositorio digital.

En la evaluación de las funcionalidades se puede notar que DSpace sigue sobresaliendo en más aspectos que su competencia, lo que reafirma con más contundencia la conclusión que me llevó a elegir a DSpace como la mejor herramienta de las cuatro.

Primeramente se nota que es muy eficiente por su método de indizado (y por todo lo dicho anteriormente), también sobresalen Alfresco y SharePoint. De igual manera sucede con los tipos de búsquedas y que como ya se mencionó, son derivadas de las excelentes formas de indizado.

Dentro de los rubros de importación y exportación es muy grato ver que todas nos permiten realizarlo y que incluso trabajan muy bien entre ellas, tanto de licencia libre como comercial.

El siguiente parámetro es el del manejo de estadísticas de uso y bueno, es evidente cual herramienta es la que más destacó. DSpace ofrece esta característica y también se hace mención de que esta herramienta, a Alfresco y a SharePoint Server se les puede expandir las funcionalidades desarrollando códigos de programación por cuenta propia de los usuarios

La característica de permitirnos manejar más de una colección se hace presente en las cuatro herramientas evaluadas. Otro aspecto que destaca es la personalización de la interfaz, aquí solamente las herramientas Greenstone y Alfresco ofrecen (burdamente) la posibilidad de modificar esta parte de manera nativa, sin embargo, las otras dos herramientas aunque no lo permiten hacer de forma inicial, son también personalizables en esta característica co la ayuda de programadores especializados en TI.

Sobre el manejo de usuarios, es una característica que le da a DSpace mucha ventaja con respecto a las otras tres ya que ésta herramienta permite crear múltiples tipos de usuarios y grupos (en comparación a Alfresco, Greenstone y SharePoint) desde su interfaz asignándoles también roles.

El tipo de recursos electrónicos que manejan todas las herramientas son muy variados y creo yo que prácticamente ofrecen el almacén de la gran mayoría de objetos que se pueden utilizar en un repositorio digital.

El uso de un catálogo de autoridades también es muy útil y DSpace también ofrece esta opción, esto habla de que la herramienta está muy bien estandarizada y brinda apoyo al usuario para minimizar errores que pueden derivarse de la falta de conocimiento sobre los repositorios digitales.

El hecho de poder compartir colecciones es también derivado de los buenos protocolos de comunicación que tienen las herramientas y tal como indica la tabla 8 todas brindan esta característica, con esto se puede pensar en las infinitas posibilidades que se tienen sobre el intercambio de conocimientos entre distintos repositorios de alguna zona y el alto crecimiento cultural que se puede brindar.

En conclusión, para mí el criterio de la elección de una de las herramientas de construcción de repositorios digitales se debe basar en todos los aspectos que se han estado discutiendo anteriormente, no solo es cuestión de posibilidad económica sino que se deben evaluar todos los puntos anteriores para tener un estudio más completo sobre la elección de una herramienta sobre otra, se deben considerar muchos aspectos y situaciones para poder operar un repositorio digital y en este trabajo se han presentado ventajas y desventajas de cuatro de las herramientas más modernas tanto de libre licencia como de licencia comercial reiterando que para mí la más completa es DSpace.

La realización de este trabajo también implicó el estudio de muchas áreas del conocimiento en ciencias de la computación como son: Bases de datos, Redes de Datos, Tecnologías de lenguajes de programación orientados a objetos, protocolos y estándares de tecnologías de información, entre otros y no está de más mencionar que una segunda adaptación de este estudio puede incluir material un poco más elaborado como la realización de manuales y videos acerca de la instalación, operación y mantenimiento de las mismas herramientas evaluadas.

ANEXOS

ANEXO A. PARTES MÁS ESENCIALES QUE CONFORMAN UN REPOSITORIO

En general todas las herramientas de repositorios digitales se conforman de las mismas partes, algunas herramientas pueden tener más partes que otras pero por lo general sí tienen en común algunas partes esenciales.
Por ejemplo la siguiente imagen del portal de un repositorio nos describe las partes más esenciales del mismo.

En la sección del banner se suele poner aquella información que resalta como único un repositorio de otro, imagen corporativa del sitio, algún slogan y/o logotipo empresarial si es un repositorio privado o personalizado, si es un repositorio público generalmente lleva la firma de la herramienta.

El buscador es de las herramientas de más utilidad e importancia en un repositorio ya que permite al usuario interactuar con el contenido almacenado digitalmente. Dependiendo el software de indización que tiene el repositorio será el tipo de búsqueda que permita hacerle a la información.

Las listas alfabéticas nos permiten filtrar la información de los repositorios por medio de un listado que, dependiendo la herramienta, permitirá escoger entre varios tipos como por ejemplo filtrados por autor, por materia o por fecha.

En la sección de la administración del repositorio podemos encontrar utilidades para el control entre el usuario y la herramienta, generalmente se forma por otros subsistemas que brindan más usabilidad al repositorio y que tienen que ver con la administración de éste, podemos encontrar por ejemplo la característica de podernos registrar al portal, revisar el perfil del usuario, notificaciones, ayudas, links a otros repositorios, entre otras.

La parte referente a la información del repositorio y navegación consiste en mostrar propiamente el contenido del material a consultar al usuario. Permite además navegar entre los distintos contenidos alojados.

Por último tenemos la sección de usos múltiples, este tipo de sección es muy variada entre los repositorios, o bien, no siempre contienen lo mismo, generalmente es una sección que contiene herramientas muestran al usuario información extra sobre la herramienta que está operando. Se pueden ver guías, links a otros recursos, sellos de validación digital y/o avisos al usuario.

ANEXO B. INGESTA, BÚSQUEDA Y CONSUMO EN COLECCIONES DIGITALES

El proceso de ingesta consiste en almacenar objetos a una colección digital, sin embargo, también hace referencia a la clasificación de éstos objetos con el fin de poder garantizar un buen resultado a la hora de hacerle búsquedas. Para esto se cuenta con técnicas como la asignación de metadatos descriptivos que ordenan al objeto dentro de una colección en el lugar preciso del tema al que está haciendo referencia. La ingesta de objetos se da a través de los usuarios del repositorio.

Figura 24. Ingesta en una colección digital

La búsqueda nos permite obtener la información que deseamos retirar dentro de un repositorio, una buena búsqueda está ligada a los tipos de indizado que hace el repositorio y a la buena catalogación y asignación de metadatos a los objetos.

Figura 25. Búsqueda en una colección digital

Cuando realizamos la búsqueda se obtiene un listado de registros con los posibles valores que nos llevan al objeto o recurso digital de interés. Al proceso de elección y extracción (uso) de un objeto se le conoce como consumo.

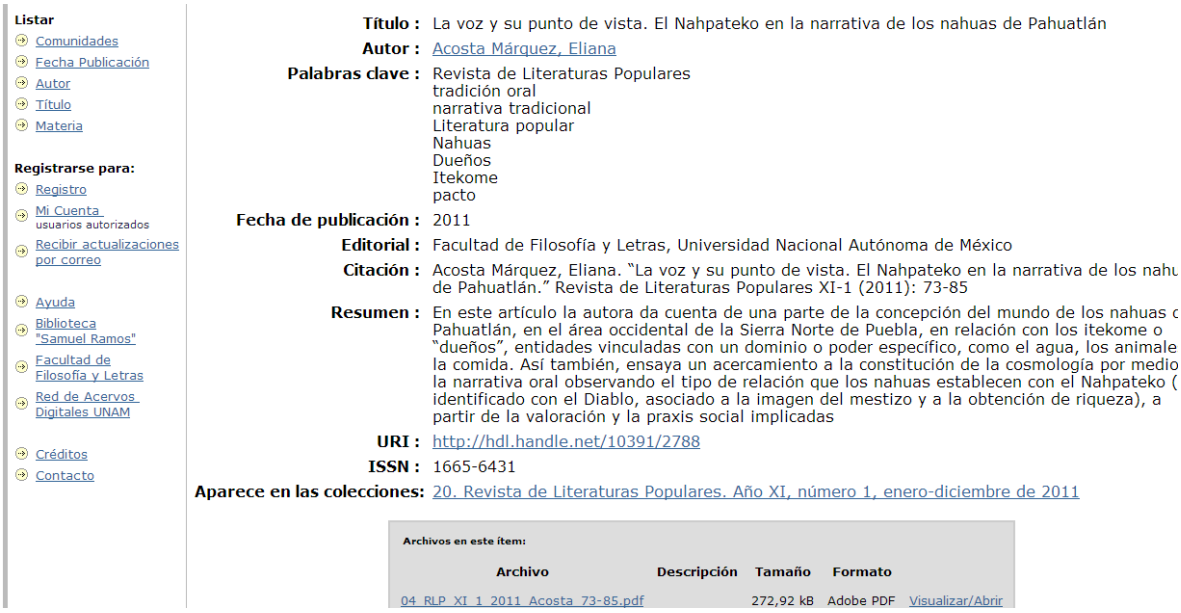


Figura 26. Consumo en una colección digital

REFERENCIAS

Alfresco Help.
Retrieved from: http://docs.alfresco.com/4.0/index.jsp [Consultado: 07/10/12].

Aggarwal, Charu C., & Yu, Philip S. (2001). On effective conceptual indexing and similarity search in text data. International Conference on Data Mining, 3-10. doi:10.1109/ICDM.2001.989492 [↑]

Basic tasks in SharePoint Server 2010 - SharePoint Server - Office.com. Retrieved from: http://office.microsoft.com/en-us/sharepoint-server-help/basic-tasks-in-sharepoint-server-2010-HA101839175.aspx
[Consultado: 08/10/12].

Bia, Alejandro., & Sánchez, Manuel. (Septiembre 2002). Desarrollo de una política de preservación digital: tecnología, planificación y perseverancia. Jornadas sobre Bibliotecas Digitales.
Retrieved from
http://mariachi.dsic.upv.es/jbidi/jbidi2002/Camera-ready/Sesion1/S1-4.pdf [↑]

Bne.es (n.d.). Diccionario de Datos PREMIS de metadatos de preservación. Biblioteca Nacional de España.
Retrieved from:
http://www.bne.es/es/Micrositios/Guias/DiccionarioPremis/Introduccion/Antecedentes/Metadatos/ [Consultado: 14/10/12]. [↑]

Botello Castillo, A. (2002). Construcción de Servicios Web con SOAP. *Revista Digital Universitaria*, Iss. 1. [↑]

Configurando Sharepoint 2010 Fast Search Server | diazantuna.es.
Retrieved from: http://www.diazantuna.es/?p=1033 [Consultado: 08/10/12].

Euán Avila, J.I., & Cordero B., L. G. (1989). *Estructuras De Datos.* México,Noriega: Limusa. [↑]
García Cárdenas, E. (2007). *Protocolos de Intercambio de Información y Lenguajes de Representación de Contenidos*. México, D.F.: Dirección General de Bibliotecas, UNAM. [Consultado: 01/10/12]. [↑]

Groove Server 2010 features and benefits - Office Servers - Office.com. Retrieved from: http://office.microsoft.com/en-us/servers/groove-server-2010-features-and-benefits-HA101810271.aspx?CTT=3
[Consultado: 08/10/12].

GUÍA DE INSTALACIÓN.
Retrieved from:
http://www.greenstone.org/manuals/gsdl2/es/html/Install_es_index.html
[Consultado: 06/10/12].

GUÍA DEL USUARIO.
Retrieved from:
http://www.greenstone.org/manuals/gsdl2/es/html/User_es_index.html
[Consultado: 06/10/12].

Hammer S., Dickmeiss A., Levanto H., Taylor M. (2005). Zebra-User´s Guide and Reference. [↑]

Hernánez Zapardiel, Ignacio José. (Diciembre 2005) Métodos y Políticas de Respaldo (backup) en Planes de Contingencia. Universidad Politécnica de Madrid, España.
Retrieved from:
www.criptored.upm.es/guiateoria/gt_m001l.htm [↑]

Home - DSpace - DuraSpace Wiki.
Retrieved from: http://wiki.duraspace.org/display/DSPACE/Home [Consultado: 07/10/12].

Huaroto, L. (2007). *El Protocolo OAI-PMH y su aplicación en el ámbito universitario*. Arequipa: Biblioteca Central - UNMSM.
Retrieved from:
http://www.slideshare.net/lhuaroto/oai-pmh-y-su-aplicacion-en-el-ambito-universitario [Consultado: 01/10/2012]. [↑]

Hudson, A., & Hudson, P. (2008). *Fedora Unleashed.* (8 ed.). Indianapolis, Ind: Sams. [↑]

Johnson, J. L., Ramírez, G. E. & Romo, M. J. H. (2000). *Bases De Datos: Modelos, Lenguaje, Diseño.* México: Oxford University Press. [↑]

Jones, M., & Beagrie N. (2001). *Preservation Management of Digital Materials*. British Library Cataloging in Publication Data. [↑]

Keefer, Alice., & Gallart, Núria. (2003). *La preservación digital y las universidades: el estado de la cuestión*. 8as Jornadas Españolas de Documentación.
URI: http://hdl.handle.net/10760/6780 [↑]

Lamarca Lapuente, M. (2011). *Hipertexto: El nuevo concepto de documento en la cultura de la imagen* (Tesis Doctoral. Universidad Complutense de Madrid). [↑]

Lara Pacheco, G. et al. (2008). *Digitalización de colecciones, Texto e imagen. Volumen 1*. [e-book] México: Dirección General de Bibliotecas UNAM, Dirección General de Servicios de Cómputo Académico UNAM, Coordinación de Universidad Abierta y Educación a Distancia UNAM.
Retrieved from:
http://www.digitalizacion.unam.mx/ [Consultado: 01/10/12]. [↑]

Library.cornell.edu (2003). *Llevando la teoría a la práctica. Tutorial de Digitalización de Imágenes - Metadatos*.
Retrieved from:
http://www.library.cornell.edu/preservation/tutorial-spanish/metadata/metadata-01.html [Consultado: 01/10/12]. [↑]

Long, L., Díaz, D. J. J. & Sánchez, G. G. (1995). *Introducción a Las Computadoras Y Al Procesamiento De Información.* (4 ed.). Naucalpan de Juárez (México) [etc.: Prentice-Hall Hispanoamericana. [↑]

López, C., & García, F. J. (2007). Los repositorios digitales en el ámbito universitario. En "La Memoria de Virtual Educa 2007. São José dos Campos - São Paulo, Brasil)". URI: http://hdl.handle.net/10366/55713 [↑]

Manual - GreenstoneWiki.
Retrieved from: http://wiki.greenstone.org/wiki/index.php/Manual [Consultado: 06/10/12].

Martínez Gallo, J. (n.d.). *El z39.50*. España: SEDIC.
Retrieved from: http://www.sedic.es/z3950.pdf
[Consultado: 01/10/12]. [↑]
McGray, A.T., & Gallagher M.E. (2001). Principles for Digital Libraries Development. *Communications of the ACM*, 44, 49-54. [↑]

Norma cubana (1985). Métodos para el análisis de documentos, determinación de su contenido y selección de los términos de indización (ISO 5963: 2000 idt). Retrieved from: http://www.sld.cu/galerias/pdf/sitios/centromed/nc_iso_5963_metodos_para_el_analisis_de_documentos,_determinacion_de_sucontenido_y_seleccion_de_terminos_de_indizacion.pdf [Consultado: 23/09/2012]. [↑]

Novedades de SharePoint Workspace 2010 - SharePoint Workspace - Office.com. Retrieved from: http://office.microsoft.com/es-es/sharepoint-workspace-help/novedades-de-sharepoint-workspace-2010-HA010288176.aspx [Consultado: 08/10/12].

Openarchives.org (2002). *Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0*. Retrieved from: www.openarchives.org/OAI/2.0/openarchivesprotocol.htm [Consultado: 01/10/12]. [↑]

Ortíz Ancona, Dante. (2010). Preservación y Conservación Digital. México: División General de Bibliotecas. [Consultado: 01/10/12]. [↑]

Ortíz Ancona, Dante. (2003). *Sistema de Transacciones Cooperativas para un Ambiente de CASE*. (Tesis de Maestría, DGEP UNAM). Retrieved from: http://132.248.9.195/pdtestdf/0325395/Index.html [↑]

Ortíz Ancona, Dante. (2007). *Software libre en la representación, búsqueda, recuperación e intercambio de información*. En CUIB UNAM (chair). Memorias de I Simposio Internacional sobre Organización del Conocimiento: Bibliotecología y Terminología. México, D.F. [↑]

Preserving our digital heritage. (October 2002). *Plan for the National Digital Information Infrastructure and Preservation Program, A collaborative Initiative of the Library of Congress*. Retrieved from: http://www.digitalpreservation.gov/documents/ndiipp_plan.pdf [↑]

Quickstart Guide - www.dspace.org. Retrieved from: http://www.dspace.org/quick-start-guide [Consultado: 07/10/12].

Reference Model for an Open Archival Information System (OAIS). (2002). *Recommendation for Space Data Systems Standards, Consultative Committee for Space Data Systems, CCSDS 650.0 –B-1*. Retrieved from http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/Ccsds-650.0-B-1.pdf [↑]

SharePoint Workspace 2010 features and benefits - SharePoint Workspace - Office.com. Retrieved from: http://office.microsoft.com/en-us/sharepoint-workspace/sharepoint-workspace-2010-features-and-benefits-HA101807659.aspx [Consultado: 08/10/12].

Sistema de gestión de contenido (CMS) empresarial de código abierto | Alfresco. Retrieved from: http://www.alfresco.com/es [Consultado: 07/10/12].

SkunkWorks SharePoint Portal. Retrieved from: http://www.gavd.net/servers/sharepointv4/spsv4_item.aspx?top=inf&itm=1092 [Consultado: 08/10/12].

Soporte : Greenstone Digital Library Software. Retrieved from: http://www.greenstone.org/support_es [Consultado: 06/10/12].

System requirements for Groove Server 2010.
Retrieved from: http://technet.microsoft.com/en-us/library/ee681781.aspx
[Consultado: 08/10/12].

Top 10 reasons to try SharePoint Workspace 2010 - SharePoint Workspace - Office.com. Retrieved from: http://office.microsoft.com/en-us/sharepoint-workspace/top-10-reasons-to-try-sharepoint-workspace-2010-HA101631747.aspx?CTT=3 [Consultado: 08/10/12].

Use content types to manage content consistently on a site - SharePoint Foundation - Office.com. Retrieved from: http://office.microsoft.com/en-us/sharepoint-foundation-help/use-content-types-to-manage-content-consistently-on-a-site-HA010375560.aspx
[Consultado: 08/10/12].

Ventajas de SharePoint 2010. Retrieved from:
http://www.desarrolloweb.com/articulos/ventajas-sharepoint-2010-dotnet.html
[Consultado: 08/10/12].

W3.org (2001). *SOAP Version 1.2*. Retrieved from: http://www.w3.org/TR/2001/WD-soap12-20010709/ [Consultado: 01/10/12]. [↑]

Waugh, Andrew., Wilkinson, Ross., Hills, Brendan., & Dell'oro, Jon. (2000). Preserving Digital Information Forever, Digital Libraries, san Antonio TX. ACM 1-58113-231-X/00/0006. [↑]

Witten, Ian H., Moffat, A., & Timothy C.(1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*, Second Edition. USA: Morgan Kaufman Publishers, Inc. [↑]

What\'s New in Microsoft SharePoint Server 2010 - SharePoint Server - Office.com. Retrieved from: http://office.microsoft.com/en-us/sharepoint-server-help/what-s-new-in-microsoft-sharepoint-server-2010-HA010370058.aspx [Consultado: 08/10/12].

www.dspace.org. Retrieved from: http://www.dspace.org/ [Consultado: 07/10/12].

BIBLIOGRAFÍA

Abraham S., Henry F. K., S. Sudarshan, (2002). *Fundamentos de Bases de Datos.* España: McGrawHill. ISBN: 0-07-228363-7 [Consultado: 08/05/13]. [↑]

Asociación Española de Normalización y Certificación AENOR. (1991). Métodos para el análisis de documentos, determinación de su contenido y selección de los términos de indización. Madrid Retrieved from: http://docubib.uc3m.es/CURSOS/Documentos_cientificos/Normas%20y%20directrices/UNE_50121=ISO%205963.pdf [Consultado: 07/05/13]. [↑]

Atre, S., & Domínguez, R. A. C. (1988). *Técnicas De Bases De Datos: Estructuración En Diseño Y Administración.* México: Trillas. ISBN: 968-24-2643-X

Cicei.com (2003). *Tutorial y descripción técnica de TCP/IP - HTTP*. Retrieved from: http://www.cicei.com/ocon/gsi/tut_tcpip/3376c426.html [Consultado: 01/10/12].