UNIVERSIDAD IBEROAMERICANA

ESTUDIOS CON RECONOCIMIENTO DE VALIDEZ OFICIAL
POR DECRETO PRESIDENCIAL DEL 3 DE ABRIL DE 1981

MARCADO ESTRUCTURAL, XML Y RECUPERACIÓN DE INFORMACIÓN EN ARTÍCULOS DE REVISTA EN TEXTO COMPLETO

TESIS

Que para obtener el grado de

MAESTRA EN INGENIERÍA DE SISTEMAS EMPRESARIALES

Presenta:

ALMA BEATRIZ RIVERA AGUILERA

Aprobado:

Mtra. Clara López Guzmán, Directora
Dr. Alfonso Miguel Reyes, Lector
Ing. Guillermo Mallén Fullerton, Lector

México, D.F. 2004

RESUMEN

Esta investigación considera la dificultad real de recuperar información de forma precisa en grandes repositorios de documentos en texto completo. La hipótesis principal de la investigación es que:

> *Al marcar artículos de revista con el vocabulario XML adecuado los resultados de la búsqueda sobre ciertas partes marcadas del texto de los artículos serán más precisos que la misma búsqueda sobre el texto completo.*

Al comprobarse la hipótesis anterior como cierta es posible que los autores, editores y administradores de repositorios de información consideren útil el marcado estructural de los artículos de revista, y los usuarios tengan el beneficio de mejorar los niveles de precisión en sus resultados de búsqueda. Además, la generación de un vocabulario simplificado y en idioma español para el marcado de artículos de revista resulta atractiva al conocer que los vocabularios existentes son extremadamente complejos, enfocados a objetivos específicos y con etiquetas en idioma inglés. Por otro lado, la revisión metodológica para la generación de vocabularios XML propios de bibliotecas digitales resulta un ejercicio de singular interés.

Este estudio es descriptivo ya que reporta cómo se desarrolla un vocabulario XML para textos completos de artículos de revistas, la metodología de marcado y finalmente la medición, a nivel exploratorio, de los resultados de dicho marcado sobre la precisión en la recuperación de la información. Para el desarrollo del vocabulario se revisaron varias opciones metodológicas y se seleccionó por la propuesta de Maler, E. y El Andaloussi, J.

El diseño de investigación es del tipo cuasi experimental, con grupos evaluadores especificados con anterioridad y conformados por especialistas en educación de la comunidad de la UIA. La colección de documentos marcados es un grupo de 29 artículos de revistas mexicanas sobre educación, filtrados por un buscador en la opción de texto completo, sobre el que se aplicaron diversas estrategias de búsqueda establecidas con ocho expertos en educación. Sobre el conjunto de artículos recuperados se hizo una valoración de relevancia uno a uno. Finalmente se identificaron los sets generados si la estrategia se aplicaba solo a ciertas etiquetas del texto y se retomó la valoración que el experto dio a los artículos, para calcular los porcentajes de precisión en ambos casos de búsqueda: sobre texto completo y sobre etiquetas.

La hipótesis fue encontrada verdadera en el contexto del estudio exploratorio ya que al llevar a cabo búsquedas sobre textos marcados la precisión fue del 62.84 % y al buscar en texto completo la precisión fue del 40.72 %.

Los resultados de esta investigación pueden dar origen a estudios con muestras significativas y resultados probatorios; así como a esfuerzos interinstitucionales para el desarrollo de vocabularios consensuados potenciales de ser considerados espacios de nombres de XML, que si se implementan por una comunidad de usuarios apoyarán la interoperabilidad entre los sistemas.

Agradecimientos

Esta investigación nació de mi solicitud a la Mtra. Clara López de la UNAM-DGSCA (Universidad Nacional Autónoma de México, Dirección General de Servicios de Cómputo Académico) de un tema de tesis que encauzara mi experiencia de casi 20 años en sistemas relacionados con bibliotecas universitarias y mi interés por las nacientes bibliotecas digitales; el desarrollo de la tesis se dio bajo su constante apoyo y supervisión. El Dr. Alfonso Miguel y el Ing. Guillermo Mallén fueron maestros excepcionales del programa académico que finalizó con este trabajo, y lectores exigentes y generosos, esa combinación que sólo logran los buenos maestros. El Mtro. Fernando Álvarez, actual Director de la Biblioteca Francisco Xavier Clavigero, plantó en mí el interés por los lenguajes de marcado y apoyó el desarrollo de este trabajo durante 2003, mi año sabático como académica de la Universidad Iberoamericana. Las ingenieras Rosita García y Angélica Corral, de la UNAM-DGSCA me iluminaron con su conocimiento y experiencia sobre Harvest y XML; y la Mtra. Isabel Gallina del mismo centro de trabajo me facilitó el acceso a parte de los archivos de artículos de revista que se marcaron. El Ing. Héctor Masayuki Hernández Hayashi y el Mat. José Juan Téllez de la Universidad Iberoamericana, Campus Ciudad de México me proporcionaron gran apoyo en aspectos administrativos de Linux. La participación de los académicos del Departamento de Educación y del Centro de Formación Valoral en las pruebas de búsqueda y evaluación de la información fue fundamental para este trabajo.

Gracias al Internet pude contar con valiosos comentarios del Dr. Ricardo Baeza-Yates de la Universidad de Chile, el Ing. Rick Beaubien de la Universidad de Berkeley, el Mtro. Alejandro Bía de la Universidad de Alicante, Los Drs. Norbert Fuhr y Saadia Malik de la Universidad de Duisburg-Essen, la Dra. Grete Pasch de la Universidad Francisco Marroquín de Guatemala, el Ing. Oliver Pesch de EBSCO Information Service y la Mtra. Pippa Smart del INASP (International Netwwork for the Availability of Scientific Publications).

Agradezo a mi esposo Blas, por su gran paciencia y apoyo en el curso de esta investigación y especialmente por la revisión del manuscrito; y a mi hijo Ignacio su comprensión y curiosidad por las horas que mamá se pasó leyendo y ante la pantalla de la computadora y especialmente por la nubes de la Figura 4.3.

INDICE

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

1.1 Antecedentes de la investigación

Los repositorios electrónicos de datos y documentos y el acceso remoto a los mismos a través del Internet han impactado a las organizaciones que almacenan registros de información. La gran mayoría de las instituciones han experimentado una transición de archivos en papel hacia medio digital y su consecuente publicación en la red, ya sea limitada al acceso interno o abierta al mundo. El aprovechamiento del contenido de innumerables documentos que almacenan datos en forma de texto, imagen, sonido o video, es ahora un reto tan vigente como siempre; con la diferencia de que los acervos analógicos no tenían las posibilidades de explotación y acceso que prometen los acervos digitales.

Las instituciones de educación superior son un ejemplo de las organizaciones que están modelando las tareas universitarias de enseñanza, aprendizaje, investigación y administración universitaria, tomando en cuenta las posibilidades que ofrece la tecnología digital. Dentro de las universidades un caso particular son las bibliotecas, que como repositorios de información en apoyo a la docencia y la investigación, se han visto ante la necesidad de adaptar sus colecciones y servicios tradicionales a los formatos digitales.

Esta investigación está enfocada al ámbito de la recuperación de información en las colecciones de bibliotecas digitales (ver definición en apartado 1.6) universitarias, específicamente a la necesidad de estructurar y marcar los artículos de revistas académicas de forma que la recuperación de información a través de buscadores proporcione mejores resultados a quien busca. Esta propuesta, aunque enfocada a artículos de revista puede ser extrapolada a otros documentos

de texto en bibliotecas digitales o en acervos digitales en general: manuales, reportes, informes, etc. que posean un cierto grado de uniformidad en su estructura. La homogeneidad en las estructuras de archivos, conjuntamente con los datos catalográficos o descriptivos sobre el contenido de los mismos, facilitan tanto la recuperación como la interoperabilidad entre repositorios digitales.

Al igual que las colecciones en papel en una biblioteca tradicional, la generación y administración de acervos electrónicos tiene objetivos entre los que se destacan los de tipo educativo y cultural. Dowler, L. (1997) en su libro “Gateways to Knowledge” (Portales de conocimiento) presenta reflexiones y datos sobre el impacto de las tecnologías en la enseñanza, el aprendizaje, la investigación y el papel de las bibliotecas como portales hacia el conocimiento; se desearía al menos como portales eficientes hacia la recuperación de la información y los contenidos. La educación a distancia en línea, por ejemplo, no se puede concebir sin el complemento de la biblioteca digital que sirva de apoyo a quienes desean seguir cursos por Internet disponiendo de material de consulta para sus tareas e investigaciones.

En los inicios de la preocupación sobre el tema de bibliotecas digitales en México, Malo Álvarez, S. y Fortes Besprovani, M.[1] (1999) propusieron la generación de colecciones y servicios digitales enfocada a las necesidades de la sociedad mexicana en la era de la información, a través de una red de bibliotecas digitales. Esta propuesta no se concretó; sin embargo, una diversidad de colecciones se

[1] Aunque la propuesta de estos autores solicitada por la Academia Mexicana de Ciencias hace ya 4 años no se han concretado el material sigue siendo muy valioso por su enfoque y síntesis acertada. Entre las razones que se considera incidieron en la falta de concreción de este tipo de proyectos puede vislumbrarse la falta de voluntad política y la cultura misma de los desarrollos tecnológicos en Internet que es de tipo más bien independiente y descentralizado.

han ido creando en toda la República Mexicana en especial en el ámbito de la educación superior (Sánchez Huitrón, J. A y Fernández, M.L., 2000, y Sánchez Huitrón, J. A, 2002).

La infraestructura de comunicaciones permite actualmente mayor ancho de banda y opciones como Internet 2 (Li, C. y Stone, H.S, 1999) presentan un espacio idóneo, todavía no completamente explotado para la transferencia masiva de datos.[2]

Al generarse cada vez más colecciones de documentos digitales el tema de la eficiencia en el manejo de los grandes repositorios de información en las llamadas bibliotecas digitales se ha convertido en un asunto de interés para disciplinas como la computación y la bibliotecología (Fox, E. y Urs, S. R. 2002 y Borgman, C. L. 2000). La problemática es compleja y abarca los grandes temas de almacenamiento, preservación del material digital, interoperabilidad, administración de contenidos; y por último, pero no menos importante, el problema de la recuperación de material pertinente para el usuario (Fox, E. y Urs, S.R. 2002 539-540; Witten, I.H. y Bainbridge, D. 2003; Borgman , C. L. 2000; y Lancaster, F.W. 1993).

Diversos autores (Witten, I.H. y Bainbridge, D. 2003, cap. 5; Harold, E. R. 2001: Duke, J.K. 1989; Boiko, B. 2001b) han considerado que el marcado de los textos electrónicos puede ser valioso en el contexto de los sistemas de administración de documentos digitales, ya que se caracteriza por:

1. Diversificación de las presentaciones a partir de una misma fuente de datos
2. Independencia de software de los archivos generados

[2] http://www.cudi.edu.mx

3. Legibilidad humana de los contenidos
4. Mejora de la calidad de los resultados en una búsqueda sobre este tipo de material.

Este trabajo se enfoca en el último punto y trata de conocer el impacto sobre los resultados de búsqueda del uso de lenguajes de marcado[3], especialmente XML (Extensible Markup Language, Lenguaje Extensible de Marcado), en textos de artículos de revista, tomando como evaluadores de la utilidad de los resultados de una búsqueda a profesores universitarios. El indicador de impacto que se utilizará será la llamada precisión, la cual es una medida comúnmente aceptada en el área de búsqueda y recuperación de información y que se calcula como el porcentaje de documentos recuperados a través de una búsqueda que es valuado como relevante por quien necesita la información. (Lancaster, F.W. 1979, pp. 108-120, 1983, pp. 159-186, y 1993, pp. 181-219; Witten, I.H, Moffat, A. y Bell, T.C., 1999, pp. 188-189, y Borgman, C.L., Moghdam, D. y Corbett, P.K., 1984, pp. 133-145)

El pilar de este trabajo es el XML, el cual se trata de un conjunto de reglas para definir etiquetas semánticas que descomponen un documento en partes identificables. XML es un lenguaje de metamarcado que define una sintaxis para escribir lenguajes específicos de marcado, también llamados vocabularios. (Harold, E.R. 2001 y Morrison, M. 2000)

Hoy día se reportan pocos editores y autores de textos electrónicos que utilicen el marcado XML para la generación de sus contenidos a nivel del cuerpo de sus

[3] En realidad todos los archivos digitales de texto poseen marcas para indicar como se debe procesar, el concepto de marcado procede de la época en que los impresores colocaban marcas a mano sobre los manuscritos para indicar tipografías, márgenes y otras indicaciones de diseño gráfico. Lo que hoy día se llama lenguaje de marcado se refiere a las etiquetas que estructuran textos digitales y que ponen énfasis en el contenido y/o en la forma del texto.

documentos, el uso mas común es para los llamados metadatos o datos descriptivos. Algunos buscadores de texto completo aprovechan los lenguajes de marcado y aunque existen propuestas teóricas e implementaciones, en la práctica, no se pone comúnmente a disposición del usuario la capacidad de buscar dentro de un contexto documental establecido a través de lenguajes de marcado XML, en especial porque no hay una gran cantidad de texto marcado (Luk, R. W. P. et .al. 2002).

En la medida que se mida si el marcado XML de documentos es beneficioso para la recuperación de información, se justificará que autores y editores lleven a cabo el esfuerzo de estructurar y marcar los materiales, y que los administradores de grandes volúmenes de contenido utilicen y pongan a disposición del usuario final buscadores que aprovechen el marcado.

1.2 Problema de investigación, hipótesis, preguntas

A partir de la dificultad real de recuperar información de forma precisa en grandes repositorios de documentos en texto completo, se plantea como problema de investigación definir un vocabulario XML que describa adecuadamente la estructura de artículos de revistas de tipo académico en bibliotecas digitales y recoja los contenidos significativos inmersos en el texto; de forma tal que al aprovechar este marcado a través de un buscador la precisión en los resultados de una búsqueda sea más alta que si el buscador indiza la totalidad del texto.

El objetivo principal de este trabajo, es proponer y evaluar un vocabulario XML que describa y estructure adecuadamente los contenidos de artículos de revistas académicas; dicha descripción y estructuración potenciará los archivos con las bondades del XML ya descritas en el apartado 1.1, en especial se espera mejore

la recuperación de información en términos de incremento en la precisión de los resultados. Un objetivo específico es probar una metodología adecuada para el desarrollo de dicho vocabulario, la de Maler, E. y El Andaloussi, J.[4], (1996) para el desarrollo de DTD´s o esquemas[5]; un segundo objetivo específico es considerar en el desarrollo del vocabularios los esquemas de marcado ya existentes aplicables a artículos de revista.

La hipótesis principal de la investigación es que:

> *Al marcar artículos de revista con el vocabulario XML adecuado los resultados de la búsqueda sobre ciertas partes marcadas del texto de los artículos serán más precisos que la misma búsqueda sobre el texto completo.*

La evaluación de los resultados de una búsqueda ha sido explorado desde 1955 (Borgman, C.L., et. al. 1984, p. 134) cuando Kent, et.al. propusieron una serie de medidas sobre la calidad de los resultados de una búsqueda, de las cuales, ya para 1984 sólo sobrevivían dos: precisión y relación de recuperación ("recall"). Para mayor detalle del concepto de precisión ver Apartado 3.6.

1.3 Justificación de la investigación

Durante la experiencia directa con el uso de tecnología de información en la Biblioteca Francisco Xavier Clavijero (FXC) de la Universidad Iberoamericana

[4] La metodología de desarrollo de DTD de Maler. E y El Andaloussi, J. es descrita conjuntamente con otras opciones de diseño en el Apartado 3.4 y desde las revisiones bibliográficas iniciales destacó como la más completa.

[5] Los DTD y los esquemas propios del XML son las formas de expresar las reglas de composición que un archivo XML debe cumplir, en ellos se establecen las etiquetas válidas y las características de las mismas. Los DTD´s tienen una sintaxis propia y son limitados en sus tipos de datos; en cambio, los llamados XML esquemas o XSD esquemas tienen una sintaxis similar a la de cualquier archivo XML y permiten una mayor especificación a nivel de datos contenidos por una etiqueta (Ahmed, et.al, 2001)

(UIA), la cual incluyó el Programa de Biblioteca Digital, se ha encontrado el problema de la recuperación de información pertinente en los grandes repositorios de datos.

El tema es considerado relevante al combinar elementos de ciencias de la información y ciencias de la computación, ambas involucradas en las líneas de investigación de la recuperación de información y bibliotecas digitales; tópicos de gran interés en bibliotecas de instituciones de educación superior.

Existen actualmente vocabularios que enmarcan contenidos descriptivos y estructurales para objetos digitales[6], como el TEI (Text Encoding Innitiative)[7] y JAI (Journal Archiving and Interchange)[8] de tipo estructural; el Dublín Core[9] y GILS (Global Information Locator Service)[10] descriptivos y el METS (Metadata Encoding and Transmisión Standard)[11] enfocada a la administración global de archivos. Sin embargo, la idea de generar un vocabulario simplificado y en idioma español para el marcado de artículos de revista resulta atractiva al conocer que los vocabularios existentes son extremadamente complejos en su mayoría, enfocados a objetivos específicos y con etiquetas en idioma inglés. Por otro lado, la revisión metodológica para la generación de vocabularios propios de bibliotecas digitales resulta a su vez un ejercicio de singular interés.

Ciertamente, han pasado ya más de 5 años desde el inicio de esta inquietud académica, durante los cuales las tecnologías asociadas al XML han evolucionado

---

[6] Fox (2002) se refiere, entre otros autores a los objetos digitales como archivos digitales que conforman los colecciones de las bibliotecas digitales

[7] http://www.tei-c.org.uk/, utilizado en análisis de textos

[8] http://dtd.nlm.nih.gov/ Archiving and Interchange DTD

[9] http://uk.dublincore.org/schemas/xmls/, utilizado como metadatos en muchos bibliotecas digitales

[10] http://www.gils.net/index.html, utilizado por la biblioteca digital iberoamericana de la UNESCO/Univ. De Colima.

[11] http://www.clir.org/pubs/reports/pub87/pub87.pdf

y variedad de vocabularios se han propuesto[12]; sin embargo, se considera que el ejercicio sigue siendo válido ya que hoy día el marcado estructural que recoja contenidos inmersos en el texto disponible no se aplica en la mayoría de artículos de revista en formato digital con el objetivo de ser utilizado por buscadores[13].

Se espera probar este planteamiento con artículos de revista tomados de la colección digital ANUIES (Asociación Nacional de Universidades e Instituciones de Educación Superior) y de la revista DIDAC de la UIA, aprovechando el marcado a través de buscadores.

Finalmente, este vocabulario es una propuesta, que de resultar beneficiosa, será de interés para la comunidad que investiga y desarrolla aplicaciones de biblioteca digital, tanto en lo que respecta al vocabulario en sí como en las pautas metodológicas para la generación de vocabularios.

1.4 Metodología

Este estudio es descriptivo[14] ya que reporta cómo se desarrolla un vocabulario XML para textos completos de artículos de revistas, la metodología de marcado y finalmente la medición exploratoria de los resultados de dicho marcado sobre la precisión en la recuperación de la información.

Para el desarrollo del vocabulario se revisaron varias opciones metodológicas y se optó por la propuesta de Maler, E. y El Andaloussi, J. en 1996, la cual aunque diseñada para el diseño de DTD (Document Type Definition) para SGML

[12] Como los ya mencionados TEI, METS, Dublín Core y GILS

[13] Se conoce el caso de los artículos de revista publicados por Cervantes Virtual que utilizan el marcado TEI y han incursionado en el uso de etiquetas en idioma español y catalán; la IEEE y EBSCO con un marcado propietario. En dichos casos el marcado estructural se utiliza en la práctica con fines de despliegue y análisis lingüístico como en el caso del TEI.

[14] Los conceptos de tipo de estudio y diseño de investigación aquí mencionados corresponden a los expuestos por Hernández Sampieri, (1998) capítulos 4 y 6.

(Standard Generalized Markup Language) se adapta perfectamente a un desarrollo en XML para objetos digitales de texto completo y es extremadamente valiosa en términos de esta investigación por el detalle que ofrece y por considerar en una de sus etapas la comparación entre la propuesta por diseñar con vocabularios equivalentes.

Esta investigación se considera de tipo exploratorio ya que no se contó con el tiempo y el personal necesario para llevar a acabo un marcado masivo de documentos necesarios para obtener resultados probatorios. La colección de documentos marcados y el grupo de usuarios con los que se evaluaron los resultados son limitados. Pruebas con mayor cantidad de archivos serían alcanzables eventualmente pues los desarrollos técnicos ya están elaborados y probados en su mayoría, y el marcado es en sí un ejercicio simple aunque minucioso y que debe supervisarse para asegurar la calidad.

El diseño de investigación es del tipo cuasiexperimental, con grupos evaluadores especificados con anterioridad y conformados por especialistas en educación de la comunidad de la UIA. La colección de documentos marcados es un grupo de 29 artículos de revista sobre educación provenientes de las revistas DIDAC de la Universidad Iberoamericana (UIA), La Academia, de Instituto Politécnico Nacional (IPN) y Pedagogía de la Universidad Pedagógica Nacional (UPN)[15], filtrados por un buscador en la opción de texto completo, sobre el que se aplicarán diversas estrategias de búsqueda establecidas con 8 expertos en educación, quienes propondrán cada uno 3 temas a buscar y sus correspondientes estrategias de

[15] Las revistas La Academia y Pedagogía son parte de la colección de ANUIES y puede consultarse en: http://www.hemerodigital.unam.mx/ANUIES/

búsqueda utilizando palabras conectadas con operadores booleanos. Sobre el conjunto de artículos recuperados se hizo una valoración de relevancia uno a uno. Finalmente se identificaron los sets generados si la estrategia se aplicaba solo a ciertas etiquetas del texto y se retomó la valoración que el experto dio a los artículos para calcular los porcentajes de precisión en ambos casos de búsqueda: sobre texto completo y sobre etiquetas.

1.5 Esquema general de la tesis

Este trabajo consta de los siguientes capítulos:

1 Introducción, que describe los aspectos generales de la investigación y el punto de partida de la misma

2 Antecedentes y marco teórico, en el cual se hace una revisión de las temáticas y conceptos relacionados con el uso de lenguajes de marcado sobre textos completos para facilitar la recuperación de información.

3 Metodología, donde se establecen los pasos y referentes teóricos para llevar a cabo tanto el desarrollo del vocabulario XML como el establecimiento del cuasi experimento y la recolección de datos, referentes a la evaluación de los resultados de búsqueda.

4 Vocabulario XML para artículos de revista y su impacto en la precisión de los resultados de búsqueda, describe el vocabulario propuesto y la génesis del mismo; así como, la selección de un editor y de un buscador para llevar a cabo el marcado y las pruebas de precisión respectivamente. El capítulo también incluye la recolección y análisis de los datos a partir de las pruebas de búsqueda hechas por docentes e investigadores para determinar si la

aplicación del marcado tiene impacto en la recuperación de información en la colección de artículos de revista.

5 Conclusiones y recomendaciones, establece los resultados del estudio e identifica propuestas de implementación e investigaciones por seguir.

Bibliografía y anexos, éstos últimos consisten en un diagrama completo del vocabulario y la documentación de una selección de los elementos más importantes entre los incluidos en el vocabulario.

1.6 Definiciones

El siguiente glosario está basado, salvo donde se indique lo contrario, en lo expuesto por Morrison, M. (2000, pp. 852-854):

1. Biblioteca Digital. De acuerdo con la Digital Library Federation (DLF, Federación de Bibliotecas Digitales) “Las bibliotecas digitales son organizaciones que proveen recursos, incluyendo personal especializado, para seleccionar, estructurar, ofrecer acceso intelectual, interpretar, distribuir, preservar la integridad y asegurar la persistencia en el tiempo de las colecciones digitales en la medida en que estas estén, de forma rápida y económica, disponibles para el uso de una o varias comunidades de usuarios”[16]. Hoy día la misma DLF propone más que una definición, una arquitectura en este tipo de servicios, la cual asegure entre otras cosas, la transferencia de datos entre las bibliotecas mismas y los proveedores con el objetivo de dar una visión integral y más valiosa a los usuarios. Alrededor de este concepto conviven dos enfoques: el computacional centrado en los objetos

[16] http://www.dlf.org

digitales, la arquitectura y uso de los mismos y el bibliotecológico enfocado a como los servicios y estructuras tradicionales de la biblioteca se adaptan a las nuevas tecnologías. (Borgman, C. L. 2000, p.51).

2. DOM. (Document Object Model. Modelo de Objeto de Documento). Es una especificación del W3C. Provee una interfase para que programas escritos en Java, JavaScript y otros lenguajes manipulen documentos web en HTML o XML.
3. CSS. (Cascading Style Sheet. Hoja de Estilo de Cascada). Un mecanismo sencillo para añadir estilo a documentos HTML.
4. DCMI. (Dublin Core Metadata Element Initiative. Iniciativa de Elementos de Metadatos Núcleo de Dublín). Es una organización dedicada a promover la adopción generalizada de estándar de metadatos y el desarrollo de vocabularios de metadatos especializados para describir recursos electrónicos, que permitan el establecimiento sistemas de recuperación de información más inteligentes. La propuesta de vocabulario de este grupo fructificó en octubre de 2001 con la aprobación de la misma por ANSI y el establecimiento del correspondiente estándar: Z39.85-2001 (Ver apartado II.4.6.2)
5. DTD: (Document Type Definition. Definición de Tipo de Documento). La solución tradicional a la estructuración de documentos XML heredada del SGML. Las DTD se suelen usar para crear un modelo de

datos para los documentos XML, lo cual permite validar estos documentos.

6. Esquemas: Un esquema es la descripción o definición de la estructura (campos, elementos o etiquetas contenidos) de una base de datos o cualquier otra fuente de datos. (Ahmed, et.al. 2001 p.26).
7. Esquemas XML: Se refiere a la definición de la estructura de una clase de documentos XML, escritos o no en sintaxis de XML. Este término incluye DTD´s, el XSD del W3C (también conocido como Esquema XML) y otros esquemas XML de origen comercial tales como RELAX NG, XDR, Schematron, etc.
8. JAI/JP: (Journal Archiving and Interchange/Jornal Publishing. Almacenamiento e Intercambio de Publicaciones Periódicas/Publicación de Publicaciones Periódicas). Esquema o conjunto de etiquetas que describen artículos de ciencias médicas propuesto por la National Library of Medicine (Biblioteca Nacional de Medicina de los Estados Unidos).
9. Lenguajes de Marcado. El concepto de marcado procede de la época en que los impresores colocaban marcas a mano sobre los manuscritos para indicar tipografías, márgenes y otras indicaciones de diseño gráfico. Lo que hoy día se llama lenguaje de marcado se refiere a las etiquetas que estructuran textos digitales y que ponen énfasis en el contenido del texto y/o en la forma. Ejemplos de lenguajes de marcado son SGML, HTML y XML. (Maler, E. y El Andaloussi, J., 1996, p. 5) El primero y el último le dan la libertad al

usuario de definir su propio vocabulario o conjunto de etiquetas de marcado y están enfocados al contenido.

10. MARC: (Machine Readable Cataloging. Catalogación Legible por Computadora). Una aplicación concreta del Formato ISO 2709, basado en la descripción semántica y sintáctica de un estándar de intercambio de registros catalográficos.[17] Se considera un estándar de metadatos descriptivos.
11. Metadatos: Genéricamente datos sobre los datos. El concepto varía de una disciplina a otra, en el contexto de bibliotecas digitales se refiere a los datos que describen un objeto digital ya sea en sus características descriptivas (también llamadas catalográficas o de encabezado), estructurales o administrativas.
12. METS. (Metadata Encoding and Transmisión Standard. Estándar para la Codificación y Transmisión de Metadatos). Esquema de metadatos descriptivos y administrativos generado por la Biblioteca del Congreso de los Estados Unidos[18].
13. Precisión: llamado también el factor de pertinencia (Borgman, et.al., 1984, p.134) y relevancia del resultado (Lancaster, F.W., 1979, p.109). Es la tasa de registros relevantes (ver definición 15) recuperados en relación a todos los registros recuperados.

[17] http://www.loc.gov/marc
[18] http://www.loc.gov/standards/mets/

14. RDF: (Resource Description Framework. Marco para la Descripción de Recursos). Es un estándar de W3C que indica como construir vocabularios XML que describan los contenidos de los recursos web.
15. Relevancia: Indica la relación existente entre un documento y una estrategia de búsqueda a los ojos de un juez particular. La relevancia es subjetiva y no representa una relación precisa e invariante. (Lancaster, F.W. 1983, p.161)
16. SGML: (Standard Generalized Markup Language. Lenguaje de Marcado Normalizado Estándar). La primera tecnología importante sobre información estructurada normalizada, que fue creada como resultado de un trabajo hecho por IBM para ofrecer una metodología para formatear y mantener documentos estructurados, como los documentos legales. XML es un subconjunto simplificado de SGML que está orientado a la web.
17. TEI: (Text Encoding Initiative. Iniciativa de texto codificado). Las guías del TEI proveen medios para representar las características de los textos que se necesita identificar de forma explícita para facilitar el procesamiento de textos por programas de computadora. En particular la guía establece un conjunto de etiquetas que pueden ser insertadas en la representación electrónica de un texto, con el objetivo de marcar la estructura del documento teniendo como base características de interés lingüístico. Sin esas marcas explícitas muchas características de interés serían difíciles de localizar por medios mecánicos. (Ver apartado 2.4.6.2)

18. Vocabulario XML: El conjunto de etiquetas definido por un esquema XML de cualquier tipo.
19. W3C. (World Wide Web Consortium. Consorcio de la Telaraña Mundial). Grupo que desarrolla tecnologías de interoperabilidad (especificaciones, guías, software y herramientas) para obtener el máximo potencial de la Web.
20. XML: (Extensible Markup Language. Lenguaje de Marcado Extensible). Un subconjunto simplificado de SGML que incorpora muchas opciones de SGML, incluyendo la extensibilidad, la estructura y la validación. XML representa una nueva era en la Web, estableciendo un modo de transmitir datos estructurados.
21. XSL: (eXtensible Style Sheet. Lenguaje de Estilo eXtensible). Una tecnología de hojas de estilos que soporta la transformación de documentos XML de un tipo a otro, por ejemplo de XML a HTML; así como los estilos de los documentos XML en base a las reglas de formato estructurado. XSL ejecuta estos dos aspectos de la presentación de documentos XML a través de dos tecnologías distintas XSLT: XSL Transformation y XSLFO: Objetos de Formato XSL.

1.7 Alcances y limitaciones

El proceso del desarrollo de un vocabulario XML es un reto académico que dejará muchos benéficios en conocimiento y potencialmente puede llegar a convertirse en un punto de partida para la discusión y posteriores decisiones a nivel de

estándares de lenguaje de marcado tanto en el ámbito de las bibliotecas digitales tanto mexicanas como extranjeras.

Limitaciones de tiempo y de presupuesto redujeron la cantidad de documentos a marcar, así como las visitas de evaluación. Al tratarse de una investigación exploratoria los resultados son iniciales y reportan exclusivamente una tendencia.

En esta investigación no se abordará en detalle el problema de la operación del marcado mismo de los documentos, el cual es un reto en el proceso de edición digital del cual se está consciente. Los marcados automáticos, la cultura del marcado por parte de los autores y el desarrollo de editores XML realmente fáciles de usar ameritan una investigación y desarrollo independientes a este trabajo.

Este primer capítulo tuvo como objetivo establecer las bases para el trabajo desarrollado en el resto de la tesis; se revisaron los antecedentes de la investigación, se planteó un problema y una hipótesis que serán contrastadas con los datos recolectados a través de un estudio exploratorio del tipo cuasiexperimental, y se reportaron los alcances y limitaciones del estudio. Con este marco referencial se procede en los siguientes capítulos a una descripción detallada de la investigación.

2. ANTECEDENTES Y MARCO TEÓRICO

Esta investigación se enmarca en la convergencia de varios temas: bibliotecas digitales, búsqueda y recuperación de información, publicación digital y lenguajes de marcado. En este capítulo se revisarán conceptos, modelos y experiencias relacionadas con estos temas, en especial los pertinentes al problema del que se ocupa este trabajo: definir un vocabulario XML que marque adecuadamente la estructura y los contenidos textuales relevantes inmersos en artículos de revistas de tipo académico en formato digital, de manera tal que si un buscador indiza los elementos marcados, la precisión en los resultados de una búsqueda sea más exacta que si el buscador procesa la totalidad del texto[1].

2.1 Biblioteca digital

Desde la segunda mitad de la década de los 90´s hasta hoy día se ha escrito y discutido sobre biblioteca digital en distintos foros y por distintos especialistas en todo el mundo (Borgman, C.L. 2000; Witten, I y Bainbridge, D, 2003, Lesk, M. 1997 y Arms, W.Y. 2000, Fox, D. y Urs, S.R, 2002, por mencionar los ya clásicos). En el capítulo anterior (apartados 1.1 y 1.6) se revisaron varios conceptos fundamentales para este trabajo, entre ellos el concepto de biblioteca digital que pasó de enfocarse en colecciones y servicios en los 90´s a la interoperabilidad en el nuevo milenio.

---

[1] Este capítulo es resultado de una revisión bibliográfica de 1998 a 2003 e incluye libros, artículos de revista, publicaciones en la web, asistencia a conferencias, y contactos con individuos identificados como expertos en el tema a través de conversaciones personales, telefónicas y por correo electrónico. Todas las fuentes, tanto bibliográficas como personales, están listadas en el apartado Bibliografía.

Una de las características de los desarrollos de bibliotecas digitales es que profesionales de diversas disciplinas colaboran en su creación y mantenimiento, y por lo tanto han surgido diversas visiones sobre el concepto[2], al respecto Borgman, C.L. (2000, p. 51)[3], plantea que hay dos escuelas con diferentes aproximaciones:

- Comunidad de **investigadores**, en especial los que tienen formación computacional: enfocada a las colecciones y su contenido, por ejemplo, objetos digitales, interoperabilidad y búsqueda.
- Comunidad de **profesionales**, en especial bibliotecarios: enfocados a los servicios y a como las estructuras existentes se adaptan a la nueva tecnología.

Los enfoques actuales cada vez se juntan más y se observan desarrollos y propuestas con visión multidisciplinaria; lo anterior genera aportaciones valiosas en términos de organización y mejora en la entrega de recursos de información al usuario final. Borgman, C.L. (Idem, p.52) propone también una definición amplia y considera a las "bibliotecas digitales como una extensión, mejora e integración de sistemas de recuperación de información, conectados y accesibles a través de la infraestructura global de información" y señala la importancia de hacer un esfuerzo de definición para facilitar el desarrollo de la teoría, la investigación y la práctica en esta área. Witten, I. y Bainbridge, D. (2003, p.6) por su parte logran una

[2] El Mtro. Juan Voutssas durante su intervención en el Foro Interfases 2000, en la Universidad de Colima, comentaba que las bibliotecas digitales tienen diversidad de facetas de acuerdo a la problemática que se enfoque y la disciplina de los involucrados.

[3] Borgman es una autora de formación bibliotecológica, muy citada por autores de la disciplina del cómputo y ha participado activamente en las conferencias JCDL (Conferencia conjunta sobre el tema de bibliotecas digitales organizada por la ACM y la IEEE).

interesante síntesis en su concepción de biblioteca digital como “una colección enfocada a objetos digitales, incluyendo texto, video y audio, conjuntamente con los métodos de acceso y recuperación, selección, organización y mantenimiento de la colección”

El compendio de Fox, E. y Urs, S.R. incluido en el Annual Review of Information Science and Technology del 2002, del que se ha tomado la Figura 2.1, posee un enfoque multidisciplinario[4] y revisa la experiencia estadounidense sobre bibliotecas digitales con respecto a las colecciones, los servicios, los aspectos sociales, económicos y legales involucrados; provee ejemplos de aplicaciones y proporciona una riquísima bibliografía sobre el tema.

Bibliotecas Digitales: Cronología

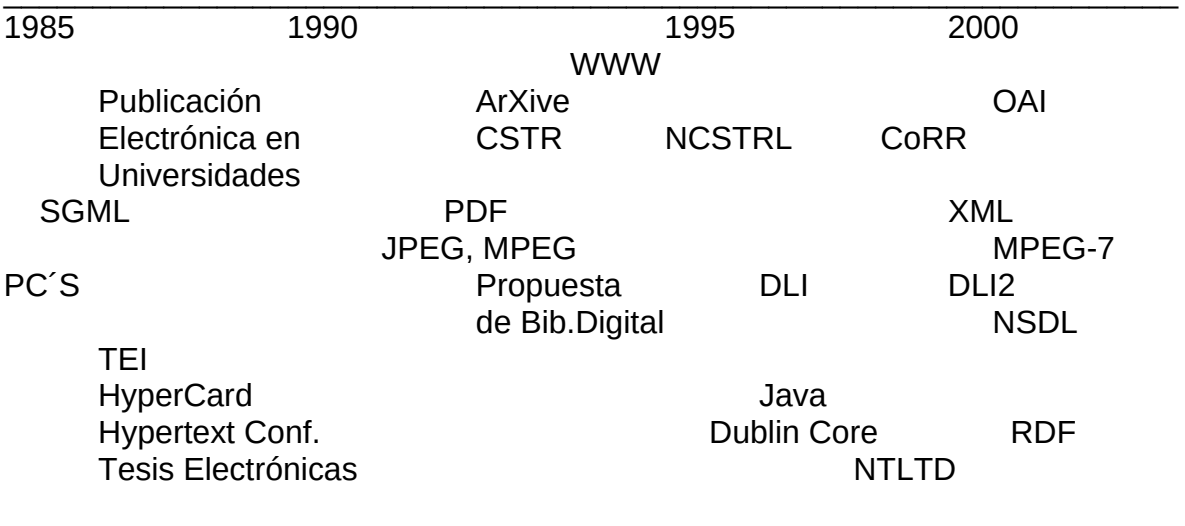


Figura 2.1. Cronología de tecnologías relacionadas con bibliotecas digitales, tomado de Fox, E. y Urs, S.R., 2002, p.508.

La figura 2.1 muestra la cronología del desarrollo de las bibliotecas digitales en la que encontramos en un primer nivel (de arriba hacia abajo) la web, le siguen los

[4] Fox proviene del área de sistemas y Urs de bibliotecología

esfuerzos de creación e interoperabilidad de archivos, y en tercer nivel encontramos los estándares. La cuarta posición la tienen los apoyos de la National Science Foundation y en el último nivel se registran las aplicaciones concretas en relación a las bibliotecas digitales.

Fox, E. y Urs, S.R. (2002, 555–556 p.) consideran que el fenómeno de las bibliotecas digitales será cada vez más común y se consolidará en colecciones y servicios. Revisada la experiencia hasta 2002 estos autores plantean como retos a futuro:

1. Una teoría unificada e incluyente en el campo de las bibliotecas digitales
2. Una metodología clara para especificar, desarrollar y mejorar bibliotecas digitales para comunidades particulares
3. Una administración de las biblioteca digitales con atención a:
    a) Balancear lo económico, social y legal
    b) Ajustarse a las ventajas de la tecnología y los estándares
    c) Considerar el ciclo de la información completo
    d) Adecuarse a los cambios de contexto de los interesados
    e) Cubrir el más amplio rango de tipos y formas posibles de contenido

Ciertamente los avances norteamericanos en este tema son importantes y se han reportado en conferencias como Digital Libraries de la ACM (Association for Computer Machinery) y de la IEEE[5](Institute of Electrical and Electronics Engineers), que a partir de 2000 se llevan a cabo como una sola conferencia, la

reunión de la American Society for Information Science and Technology.[6], el Simposio Internacional de Tesis y Disertaciones Electrónicas[7], etc., pero debe considerarse que estos desarrollos se han dado, en el mundo entero.

Las novedades sobre biblioteca digital se han publicado inicialmente en revistas[8] y hoy día aparecen cada vez más y mejores libros sobre el tema; uno de los medios en donde los avances ha sido comunicado con prontitud son las conferencias como la European Conference on Research and Advance Technology for Digital Libraries[9] cuyos temas del 2001 (Constantinopoulos, P. y Solverg, T. , 2002) incluyeron:

- Digitalización
- Interpretación y anotación de documentos
- Administración del conocimiento
- Modelos y metamodelos de datos.
- Integración y comunidades de usuarios
- Recuperación y filtro de información
- Bibliotecas digitales de multimedia y multilingualidad.

El tema de esta tesis, la recuperación de información con mayor precisión a través del uso de marcado XML está incluido en el punto sexto de los ya mencionados.

---

[5] http://www.jcdl2004.org
[6] http://www.asis.org
[7] http://www.uky.edu/ETD/ETD2004/
[8] D-Lib Magazine http://www.dlib.org
[9] http://www.ecdl2003.org

Como resultado de la globalización se tiene que en las conferencias hay presencia mundial, y personalidades como Edward Fox o Ian Witten[10] pueden encontrarse en reuniones efectuadas en diferentes continentes.

Nueva Zelanda ha sido una región que ha generado propuestas como las reportadas por Witten, I y Bainbridge, D. (2003), la cuales incluyen tanto diseño como software gratuito, y están englobadas en el concepto llamado Greenstone Digital Library, cuyo uso está promoviendo UNESCO (Organismo de la Naciones Unidas para el fomento de la Educación, la Ciencia y la Cultura) en los países en desarrollo y particularmente en Latinoamérica[11]. Los mencionados autores, en su excelente, ameno y completo manual "How to build a digital library" reportan con un enfoque más universal que los autores norteamericanos las enormes posibilidades de las bibliotecas digitales en países en desarrollo:

- Diseminar información comunitaria
- Ayuda en caso de desastres
- Preservación de la cultura indígena
- Producción de información local

Todo ello adaptado a la realidad de las limitaciones tecnológicas de las regiones en vías de desarrollo.

[10] Witten dictó una conferencia en la UDLA, campus Cholula, Puebla durante 2001. Fox fue ponente invitado en la misma institución durante la reunión del Grupo Amigos 2002 y en la II Conferencia Internacional sobre Bibliotecas Universitarias en la UNAM, octubre 2003.
[11] El Mtro. Claudio Menezes de la Oficina Regional de Ciencia para América Latina, UNESCO promocionó el software Greenstone durante la conferencia II Conferencia Internacional sobre Bibliotecas Universitarias en la UNAM, octubre 2003.

El caso chino reportado por Li, G. y Huang, M.B. (2001), muestra como el tema de bibliotecas digitales ha tenido muchos productos en el oriente en términos de publicaciones e implementaciones. Dado que, ciertamente, no todo es positivo en este tipo de proyectos, los autores reportan entre las dificultades encontradas en China la carencia de:

- Financiamiento
- Infraestructura
- Recursos en su propio idioma
- Impacto social
- Cooperación internacional.

Una experiencia de biblioteca digital cercana a Latinoamérica es el sitio español llamado Cervantes Virtual[12], el cual se ha presentado en los foros más importantes a nivel mundial y sus colaboradores han desarrollado aplicaciones interesantes en general, y en especial en la línea del análisis de textos para lo cual han utilizado el vocabulario de marcado TEI[13].

Diversos países de América Latina y África también han reportando sus experiencias en la generación de colecciones y servicios digitales, y han colaborado con proyectos como los de UNESCO sobre la Memoria del Mundo[14].

[12] http://www.cervantesvirtual.es

[13] La Universidad Iberoamericana tiene un convenio de colaboración con este proyecto desde 2001. Si bien se han enviado contenidos a ser publicados por Cervantes Virtual, lamentablemente hasta ahora no se ha logrado un intercambio tecnológico efectivo, el cual a fin de cuentas no parece ser uno de los objetivos de este proyecto de biblioteca digital con un enfoque curiosamente centralizado.

[14] http://www.unesco.org

En México han habido avances importantes como los reportados por Sánchez, A. y Fernández, M.L. en el 2000 que incluyen desarrollos de aplicaciones y colecciones digitales en instituciones de educación superior como las que aquí se enumeran sin afán de ser exhaustivos:

- Universidad de las Américas (UDLA-Cholula, Puebla)[15]
- Tecnológico de Monterrey (ITESM-Monterrey)[16]
- Universidad de Colima[17]
- Universidad Autónoma de México (UNAM-DGSCA)[18]
- Instituto Politécnico Nacional (IPN)[19].

Tanto la UDLA como la UNAM-DGSCA han utilizado lenguaje de marcado en algunas de sus publicaciones electrónicas. La UDLA ha promovido en el país el uso del estándar de interoperabilidad llamado OAI (Open Archive Innitiative) para facilitar la búsqueda en diversos repositorios de metadatos de objetos digitales y la creación de colecciones de tesis digitales. Colima ha desarrollado un modelo y metodología correspondiente a biblioteca digital en conjunto con la UNESCO. El ITESM apoyado con fondos CONACYT ha desarrollado el software libre Phronesis para administración de repositorios de archivos digitales. La UNAM-DGSCA ha desarrollado gran cantidad de colecciones digitales y ha explorado el aprovechamiento del marcado XML a través de diversos despliegues y buscadores. El IPN a través del Centro de Investigación en Computación (CIC),

[15] http://biblio.pue.udlap.mx/digital/desarrollo.html
[16] http://copernico.mty.itesm.mx/~tempo/Proyectos/
[17] http://bdigital.ucol.mx/Menu.htm
[18] http://www.bibliodgsca.unam.mx/

Coordinación de Investigación en Inteligencia Artificial, Laboratorio de Tecnologías de Lenguaje Natural, ha desarrollado propuestas de sistemas sofisticados de recuperación de información y software de administración de biblioteca digital y fue una de las primeras instituciones en México en iniciar colecciones de tesis digitales.

Además de las ya mencionadas existen otras instituciones mexicanas que han desarrollado colecciones y servicios digitales tales como la Universidad Iberoamericana, Ciudad de México con materiales antiguos y documentos institucionales, y los servicios de consulta por e-mail y chat; y el Colegio de México también con materiales antiguos. Cada vez es más común que las universidades y en especial sus bibliotecas produzcan colecciones y enriquezcan sus servicios a través de la tecnología digital.

2.2 Colecciones digitales

Las colecciones digitales son condición necesarias para la existencia de las bibliotecas digitales, diversos autores han propuesto procedimientos para construir este tipo de colecciones (Sitts, M.K. 2000; Kenney, A.R. y Rieger O.Y.1999 y Witten, I. y Bainbridge, D. 2003). Algunos elementos a tomar en cuenta para generar acervos digitales son los siguientes (Chapman, S. 1999, p.30):

1. Selección de los materiales
2. Consideraciones sobre derechos de autor
3. Preparación física del material a digitalizar si se encuentra en otro formato.
4. Asignación de metadatos

[19] http://www.cic.ipn.mx/laboratorios/coord3.htm

5. Producción digital
    a. Escaneo
    b. OCR[20]
    c. Asociación de metadatos a imágenes y archivos
    d. Retoque
    e. de imágenes
    f. Creación de versiones
6. Revisión de calidad
7. Marcado estructural en XML
8. Administración de archivos
9. Integración metadatos-imagen
10. Generación de versiones para el usuario final
11. Promoción.

Las colecciones digitales están conformadas por archivos binarios y pueden tener variedad de composiciones: texto ASCII, texto Unicode[21], imágenes en GIF, JPEG, TIFF, sonidos, videos, etc.; algunos integran varios de ellos a la vez en la llamada multimedia.

Este tipo de colecciones presentan la necesidad de administrarse correctamente para facilitar su almacenamiento y entrega a los usuarios (Rivera Aguilera, 2003). Boiko, B. (2001b) a través de su concepto de administración de contenido propone sistemas automatizados que apoyen la selección, la administración y la

[20] Optical Character Recognition. Reconocimiento Óptico de caracteres. Un proceso de conversión de imágenes a texto, existen variedad de software que lo realizan, uno de los más comunes es OmniPage.

publicación de objetos digitales. En el área de administración de contenido Boiko integra el concepto de metadatos y marcado estructural como indispensables para una mejor recuperación de la información.

La creación de colecciones digitales no requiere exclusivamente claridad sobre los elementos de tipo técnico, para su generación se necesitan habilidades administrativas tales como el establecimiento y documentación de procesos adecuados, dependiendo del tipo de material que las conformará. A partir de las experiencias reportadas en los últimos ocho años, puede indentificarse también la necesidad de voluntad política de las instituciones para apoyar este tipo de proyectos y un elemento totalmente humano: el liderazgo[22], la falta del cual ha incidido en no pocos fracasos en proyectos de generación de colecciones biblioteca digital.

Un tema que no puede dejar de mencionarse a nivel de colecciones digitales es el de derechos de autor. Muchos proyectos se han visto truncados debido al desconocimiento y la falta de claridad en la legislación, y de una política institucional sobre el tema. Un principio general al respecto, es que si un material tiene derechos vigentes, es decir, si los derechos patrimoniales de publicación no han prescrito[23], la única manera de publicarlo en formato digital es obteniendo los derechos explícitamente por parte de los titulares, en caso contrario la

---

[21] Unicode es un conjunto de caracteres (character set) que ofrece codificar todas las escrituras del mundo y al cual la mayoría de proveedores de software (Microsoft, Apple, IBM, Sun y muchos otros) están utilizando, y es el conjunto de caracteres por defecto para XML.

[22] Liderazgo que según Edward Fox en su participación en la II Reunión Internacional de Bibliotecas Universitarias, UNAM, octubre 2003 es el elemento más importante en este tipo de proyectos.

[23] Si no han pasado 100 desde la fecha de divulgación o de la muerte del autor (ver la Ley Federal de Derechos de Autor de 1996, reformas en 1997 y 2003).
http://www.impi.gob.mx/web/docs/marco_j/3w002100.htm

digitalización o la publicación del material ya digitalizado, no debe llevarse a cabo. Sin embargo algunas instituciones se han arriesgado a subir a la red o permitir subir a la red por miembros de su comunidad, materiales sin permiso explícito de los autores en el entendido de que el uso educativo es lo que se entiende como uso válido o "fair use".

Habiendo establecido el punto anterior pareciera que lo único fácil de publicar en relación al tema de derechos de autor es material antiguo[24], lo cual no siempre es la opción más fácil debido a la complejidad en la manipulación física de este tipo de material. La polémica de los derechos de autor en relación a las bibliotecas digitales ha sido ampliamente discutido en los últimos años (Universidad de Guadalajara, 2001); sin embargo todavía se presenta con poca claridad, sólo la adecuación de las leyes y las experiencias concretas irán modelando el fenómeno.

Un importante avance en el manejo de los derechos de autor en el ciberespacio son los llamados "Tratados Internet" que se refieren a los Tratados de la OMPI sobre Derecho de Autor (WCT) y sobre Interpretación o Ejecución y Fonogramas (WPPT)[25], cuyo objetivo principal es generar confianza en los creadores para crear, distribuir y controlar la utilización de sus obras en el entorno digital.

En conclusión las colecciones digitales son fundamentales para la biblioteca digital, pero su creación y administración es compleja y debe diseñarse

[24] El material antiguo tiene su propia problemática asociada a su digitalización y que está relacionada con aspectos de preservación en la cual no ahondaremos aquí, pero que puede ampliarse en los textos de Sitts. M.K (2000) y Kenney, A.R. y Rieger O.Y. (1999).

[25] http://www.ompi.org/treaties/ip/wppt/index.html y http://www.ompi.org/treaties/ip/wct/index.html

adecuadamente si se persigue la adecuada explotación de los acervos electrónicos.

2.3 Documentos, catalogación y metadatos.

En este apartado se presentan los conceptos básicos en relación a la administración de los documentos en general y a los de tipo digital en particular. Elementos claves para lograr la explotación racional de los documentos de cualquier tipo son:

- La **catalogación** como una tarea necesaria para facilitar el acceso a la información contenida en los documentos y
- Los **metadatos** como elementos facilitadores a la recuperación de información en ambientes digitales.

2.3.1 Documentos

Como ya se ha mencionado las colecciones de las bibliotecas digitales se componen de objetos digitales, es decir archivos binarios. Si consideramos que un documento (Svenonious, E. 2000, p.7) es un concretización de información en un particular espacio y tiempo, con la salvedad del multiacceso y de la replicación que permite el medio electrónico, podríamos reconocer que un objeto digital tiene la nota esencial de todo documento: contener información. Un documento digital puede ser un texto, una imagen, un archivo de música, etc. Si bien existen cada día más documentos digitales del tipo generalizado (Witten, I. H. y Bainbridge, D., 2003) o multimedia, en esta investigación se hará énfasis en los formatos textuales.

Los documentos usualmente tienen por objetivo albergar información y permitir dar seguimiento a la misma (Maler, E. y El Andaloussi J., 1996, p.38). El ser humano interactúa con los documentos a través de varias acciones:

1. Creación y modificación
2. Administración, almacenamiento y archivo
3. Uso

Para lograr una interacción eficiente con los documentos es necesario conocer a fondo los documentos en sí y las aplicaciones que permiten los tres puntos ya señalados (por ejemplo programas editores, administradores de documentos y buscadores).

Ya sea que a los elementos de una colección digital se les llame objetos o documentos, es claro que pueden describirse y estructurarse. Boiko, B. (2001b, p.11) señala 2 elementos fundamentales del contenido de los objetos digitales, los cuales nos permiten un mejor entendimiento de los mismos, a saber: formato y estructura.

1. **Formato**:

   Se refiere a la forma en que se codifica la información para que una computadora pueda leerla, en general lo que leen los equipos son 0´s y 1´s, es decir código binario, el cual puede recibirse con diferentes composiciones o formatos de archivo. El formato varía dependiendo de los objetivos y no siempre es fácil trasladar código binario de un formato de archivo a otro. Por ejemplo en imágenes para publicación en Web tenemos al JPEG y al GIF como estándares de código binario; sin embargo, para publicación impresa son más bien comunes los formatos EPS y TIFF (Boiko, B., 2001b, p. 13). Cada

estándar tiene su manera de representar el código binario, por ejemplo los gráficos vectoriales se almacenan como ecuaciones con valores que al ser calculadas permiten que se vean las formas de las imágenes.

En el caso del texto, los conjuntos de caracteres ASCII o UNICODE se consideran estándares, y sobre ellos un arreglo que lo convierte en .DOC o HTML, es también un formato muy común.

2. Estructura

Trata de cómo se organiza la información al vaciarla en un objeto digital (Boiko, B., 2001b, p. 21) y puede categorizarse por tipo de la siguiente forma (pp. 26-28):

A. *Estructura por División*: está en relación a la división del contenido en piezas utilizables, pieza puede referirse aquí desde una palabra hasta el URL de un sitio web
    a. Segmentos[26] [de una colección digital]: artículos, folletos, cartas, mails, imágenes, etc.
    b. Elementos dentro de un segmento: título, resumen, cuerpo, párrafo, texto en negrita, nota al pie, menú de opciones, etc.

B. *Estructura de Acceso*: la necesaria para acceder el contenido
    a. Jerarquías, Tablas de contenido
    b. Índices
    c. Referencias cruzadas o ligas
    d. Secuencias de vista o “browsing”

[26] El término segmento no parece muy afortunado pues puede dar lugar a confusión, sin embargo es el que usa Boiko, sugiero que se tome como segmento o pieza de una colección digital.

C. *Estructura Administrativas*: Atributos que permiten encontrar y administrar el componente de contenido. Autor, fecha de creación, número de versión, estado de revisión.

D. *Estructura inclusiva*: Qué componentes incluyen otros. Referencias a imágenes por ejemplo.

Boiko, B., (p. 29) utiliza el concepto de Arquitecto de Contenido o Metator como el individuo que divide información homogénea y la marca para tener acceso y administrar el contenido. Este profesional debe crear jerarquías, índices, estructuras de referencias cruzadas y secuencias.

El conocimiento de las estructuras y formatos documentales nos permitirá administrar con mayor eficiencia los repositorios de objetos digitales.

2.3.2 Catalogación.

Tradicionalmente la descripción de un documento se ha llamado catalogación y en esencia la catalogación y la asignación de metadatos pueden considerarse con los mismos fines. Recordemos los objetivos de un catálogo propuestos en 1904 por uno de los padres de la bibliotecología: Cutter registrado en Carpenter, M. y Svenonius, E., (1989, p. 67).

1. Permitir a una persona encontrar un libro por autor, título o tema
2. Mostrar lo que una biblioteca posee por autor, tema o tipo de literatura
3. Ayudar a la selección de un libro apoyándose en datos bibliográficos y su carácter en relación a temas y género literario.

La catalogación inicia desde antes de nuestra era al registrarse las colecciones de repositorios de información (ya sea tablillas, papiros, libros, etc.) (Schmierer, H.F.,

1989) con el propósito de facilitar su almacenamiento y recuperación. Las herramientas han sido diversas, desde las listas de títulos hasta los actuales registros Dublín Core, pasando por las tarjetas mecanografiadas y los registros bibliográficos electrónicos en formato MARC. La tecnología, se puede afirmar con Schimnierer, ha tenido impacto en la forma en que se registra información sobre una colección de documentos con el fin de facilitar su recuperación.

En esencia todas las formas de catalogación han apuntado a registrar información descriptiva de los documentos, en la era de la automatización (70´s-90´s) Duke, J.K. (1989, p. 121-124) manifestaba que un registro documental podía conformarse a tres niveles: la representación del documento, la guía del documento y el texto del documento. Los niveles corresponden respectivamente al registro descriptivo que comúnmente conocemos como cita (conformada por un conjunto de campos que identifican inequívocamente un ítem), a una sinopsis del contenido y finalmente al texto completo o partes sustantivas del mismo. Duke en 1989 pronosticaba que los materiales en texto completo todavía tardarían algunas generaciones en estar disponibles, hoy día sabemos que se equivocaba pues los textos completos son un fenómeno común actualmente. El aporte de este autor es significativo ya que propone un código de catalogación que considere los tres diversos niveles de registro ya mencionados. Duke, J.K. (1989, p. 184) establece: “Debemos comenzar a pensar en revisar el formato MARC de forma que permita que un registro bibliográfico sea visto de forma estructural” en sus tres niveles ya descritos y de forma ligada. Los beneficios que este autor establece son el manejo de los diferentes niveles según solicitud del usuario, el uso de partes del material

que podría influir en una distribución de los gastos de derechos de autor y un manejo más hábil por parte de los sistemas de piezas independientes pero ligadas de un solo documento. Duke se acerca de forma muy interesante al problema planteado en esta tesis; sin embargo, no profundiza en el uso de los niveles de registro para la recuperación de la información, los considera más bien para el despliegue. Este autor consideraba a finales de los 80's que se estaba dejando que los proveedores de sistemas para bibliotecas y de documentos en formato electrónico fueran los que decidieran las estructuras y los formatos de los contenidos documentales, y consideraba que los administradores de información podría dar ideas más versadas sobre formatos y niveles descriptivos ya que eran los usuarios intermedios y finales de los contenidos. Se considera que esta investigación es un pequeño aporte en esa línea.

Desde 1989 hasta hoy día no han pasado más que 15 años y lo que Duke esperaba tomara algunas generaciones, ya está presente: los textos completos disponibles en línea. Las propuesta de Duke pueden ser un puente para el desarrollo de los metadatos estructurales adecuados a la experiencia catalográfica y apoyar propuestas como la que se pretende en esta investigación.

2.3.3 Metadatos

Cómo ya se indicó en el Capítulo 1 los metadatos son genéricamente datos sobre los datos. El concepto varía de una disciplina a otra (Rivera Aguilera, 2001)[27], en el contexto de bibliotecas digitales se refiere a los datos que describen un objeto

[27] En minería de datos puede haber diferencias y el alcance que se pretende por ejemplo en los metadatos geográficos no será de la misma intensidad que los catalográficos.

digital ya sea en sus características catalográficas o descriptivas (metadatos de encabezado o descriptivos), o estructurales (metadatos estructurales)

Es así que en el ámbito de los documentos digitales la catalogación o descripción de documentos se hermana con el concepto de metadatos para nombrar elementos que se vacían en los sistemas para facilitar la recuperación de la información.

Ahmed, K., et.al. (2001) señala a los metadatos son un elemento clave para la administración de los repositorios digitales y la utilidad que el XML y su estándar acompañante RDF tienen en la generación y aprovechamiento de los metadatos descriptivos o comúnmente llamados metadatos.[28]

En el ámbito de los documentos comunes en bibliotecas digitales se han desarrollado ya estándares en relación a los llamados metadatos descriptivos, entre ellos el Dublín Core, el formato MARC, el METS, los inmersos en el TEI y el JAI/JP; todos ellos descritos en el apartado 1.6 del capítulo uno de esta tesis.

Para llevar a cabo su función de auxiliares en la recuperación de información de grandes repositorios de documentos, los metadatos se vacían en estructuras de de lenguajes de marcado y bases de datos, las cuales revisaremos brevemente en los apartados siguientes.

2.4 Lenguajes de marcado

En los años 70 con el afán de “estructurar documentos en forma organizada” (Morrison, 2000, p. 4-5) IBM creo GML (Lenguaje de Marcado Generalizado), y posteriormente SGML (Lenguaje de Marcado Generalizado Standard), el cual

emergió en 1986 como estándar ISO. En 1989 Berners-Lee y Berglund, del CERN (Laboratorio Europeo de Física en Partículas) generaron un lenguaje de etiquetado que facilitara el intercambio documentos científicos, dicho lenguaje se adaptó a SGML y finalmente se convirtió en el conocido HTML. En febrero de 1998 un equipo al que el W3C encargó aprovechar el poder del SGML en la web creo la primera versión del XML.

Los lenguajes de marcado o anotación (como los llaman Lafuente y Garduño, 2001)[29] tienen como principal referencia al SGML, esta tecnología define más que una forma de archivar información un metalenguaje para generar aplicaciones o vocabularios de marcado específicos con una preocupación fundamentalmente estructural[30], la cual potencia tanto el despliegue de los diversos elementos que conforman el archivo, como la recuperación de sus contenidos.

Los lenguajes de marcado (Maler, E y El Andaloussi, J. 1996, p. 3) pueden ser usados por los sistemas de cómputo para:

1. Formatear a partir de una misma fuente electrónica diversas formas electrónicas y en papel.
2. Buscar información basándose en el contexto dentro de un documento
3. Usar hiperligas

---

[28] Los conceptos aquí señalados se pueden consultar en las definiciones registradas en el apartado I.6 y se ahondará en ellos en el apartados siguiente II.4

[29] El libro de Lafuente y Garduño es una aproximación publicada en México a los temas de lenguajes de marcado y su relación con la biblioteca digital, desde un punto de vista bibliotecológico. Es un aporte valioso al integrar diversos aspectos de metadatos tanto de encabezado como estructurales y en un momento muy oportuno, lamentablemente el material no acaba de articularse de forma coherente por lo que resulta más bien un collage de notas técnicas. Más lamentable aún es que no haya más publicaciones sobre el tema en español que enriquezcan la discusión.

[30] (ver fuentes de pag. 96 de Lafuente/Garduño)

4. Tratar a los documentos como una base de datos

Como ya se mencionó en el capítulo 1 diversos autores (Witten, I.H. y Bainbridge, D. 2003, cap. 5; Harold, E. R. 2001: Duke, J.K. 1989; Boiko, B. 2001b) han considerado que el marcado de los textos electrónicos puede ser valioso en el contexto de los sistemas de administración de documentos digitales, ya que se caracteriza por:

1. Diversificación de las presentaciones a partir de una misma fuente de datos
2. Independencia de software de los archivos generados
3. Legibilidad humana de los contenidos
4. Mejora de la calidad de los resultados en una búsqueda sobre este tipo de material.

Como se ha visto hay un potencial en los lenguajes de marcado para facilitar la recuperación de información, se verá a continuación el lenguaje de marcado más utilizado hoy día.

2.4.1 XML

En términos formales (Harold, p. 3) XML es un conjunto de reglas para definir etiquetas semánticas que descomponen un documento en partes identificables, es decir permite mostrarlo de forma estructural. XML es un lenguaje de metamarcado que define una sintaxis para escribir lenguajes específicos de marcado, también llamados vocabularios; es un hijo de SGML de uso relativamente simple, enfocado al WWW y vendría a ser una versión simplificada del SGML.

Los datos de un archivo del tipo XML pueden estar en ASCII o UNICODE y no dependen de software o hardware para su almacenamiento. Es un estándar

generado y aceptado por el W3C. XML refiere pues, a una sintaxis y le es indiferente la semántica; por lo que si una etiqueta se llama autor o 100 no representa nada para el lenguaje; sin embargo para los humanos que marcan directamente o revisan el archivo obviamente las etiquetas pueden tener mucho significado. (Marko, L. ,2002)

Una de las bondades del XML que señala Harold, E.R. (2001, p.175) es que este metalenguaje provee soporte completo al conjunto de caracteres de doble-byte Unicode, así como a sus representaciones más compactas, esto significa que casi cualquier escritura moderna puede ser representada a través del XML.[31]

A continuación se describen algunos conceptos asociados al lenguaje de marcado XML y el diseño de sus esquemas (DTD o XSD)

1. **Elementos**: Contenedores anidables que permiten almacenar la información de un documento (Maler, E y El Andaloussi, J., 1996, p.12). Son la característica más importante de cualquier lenguaje basado en XML, permiten encapsular datos y proveen el espacio para los atributos. Se utilizan como envoltura de los datos que se desea registrar y de los atributos inmersos que representan los metadatos del contenido del elemento mismo. (Wyke, R.A. y Watt, A., 2002, p.71).

Se define: <!ELEMENT nombre (#PCDATA)> Se codifica: <nombre>Alma Rivera Aguilera</nombre>

Figura 2.2 Ejemplo de elemento

2. **Atributos**: Pueden ser adjuntados a los elementos para describir mejor su contenido. (Maler, E y El Andaloussi, J., 1996, p.15) o de acuerdo con Wyke

[31] Para información amplia y actualizada sobre Unicode se puede consultar http://www.unicode.org

y Watt (2002, p.121) aun sus usos. En algunos casos los atributos proveen el contenido de los datos dejando la estructura a los elementos.

```
Se define:   <!ELEMENT nombre (#PCDATA)>
             <ATTLIST nombre id #REQUIRED>
Se codifica: <nombre id=”3647”>Alma Beatriz Rivera Aguilera</nombre>
```

Figura 2.3 Ejemplo atributo

3. **Entidades**: Fragmentos de contenidos de documentos, pueden ser utilizados en otros documentos múltiples veces y actualizarse fácilmente. Existen entidades externas e internas.

```
Se define:   <!ENTITY % symbol SYSTEM "HTMLsymbol.ent"> %symbol; >
             <!ENTITY % atributos-generales "tipografia CDATA #IMPLIED
                                             numero CDATA #IMPLIED">
              <!ELEMENT titulo (#PCDATA)>
              <!ATTLIST titulo %atributos-generales; >
Se codifica: <titulo tipografia=”arial 12” >Marcado Estructural</titulo>
```

Figura 2.4 Ejemplo entidades externas e internas.

4. **Comentarios**: Permiten documentar dentro de los archivos y pueden ser utilizados en cualquier parte del documento.

```
<!— entidades necesarias para el reconocimientos de caracteres en español
<!ENTITY % symbol SYSTEM "HTMLsymbol.ent"> %symbol;
<!ENTITY % lat1 SYSTEM "HTMLlat1.ent"> %lat1; -->
```

Figura 2.5 Ejemplo de comentario.

En la figura 2.6 se encuentra un ejemplo de archivo marcado tipo XML.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="articulo.xsl"?>
<!DOCTYPE articulo.1 SYSTEM "articulo1.dtd">
<articulo.1>
<! --  comentario: El elemento relaciones incluye datos sobre la revista y el ejemplar que pueden referir a entidades si se establecen estos datos como tales  -->
 <relaciones>
    <revista>DIDAC</revista>
    <ejemplar>n41, Primavera 2003</ejemplar>
  </relaciones>
```

```
  <portada>
    <autor><nombre>Georgina Amayuela</nombre>
    <institucion>Centro de Estudios de Ciencias de la Educación Enrique José Varona Universidad
                  de Camaguey, Cuba.</institucion>
    <email>gamayuela@yahoo.com</email>
    </autor>
    <titulo>COMUNICACION  EDUCATIVA EN EL CONTEXTO UNIVERSITARIO</titulo>
<! --  el elemento resumen contiene un atributo llamado idioma  -->
 <resumen idioma=español>
      <p>En el presente articulo se parte de concebir el proceso de Comunicación y Educación
         como dos procesos complejos y que están muy estrechamente relacionados. Se enuncian
         algunos aspectos esenciales que conforman una definición integral de Comunicación. Se
         valora el impacto de ambas categorías para el proceso pedagógico, donde se concibe la
         Comunicación Educativa en su enfoque procesual.</p>
    </resumen>
  </portada>
<articulo.1>
```

Figura 2.6. Ejemplo de archivo con marcado XML, los acentos han sido removidos

Con XML un individuo puede definir las etiquetas de marcado que necesite, así como establecer como se despliegan los datos a través de un navegador utilizando un archivo CSS o un XSL.

```
COMUNICACIONOFICIAL {display: block ; font-size: 16 pt; font-wight:bold}
NUMEROCOMOF {display: block ; font-size: 16 pt; font-wight: bold; text-align: right}
FECHACOMOF {display: block ; font-size: 16 pt; font-wight:bold; text-align: right}
ENTIDAD {display: block ; font-size: 16 pt; font-wight:bold; text-align: center}
TIPOINF {display: block ; font-size: 16 pt; font-wight:bold; text-align: left}
PARRAFO {display: block;font-size: 16 pt; text-align: left}
DATOSSESION {font-size: 16 pt; text-align: center}
ENCABEZADO {font-size: 16 pt; text-align:center; display}
```

Figura 2.7 Ejemplo de hoja de estilo en cascada

Un documento XML puede contener exclusivamente elementos o combinar el vaciado de los datos en elementos y atributos de los mismos. Las especificaciones de cómo será el marcado de los documentos de datos se registra, como se verá

más adelante en los archivos tipo DTD, un esquema heredado del SGML que se fue concebido para estructurar información narrativa y sigue cumpliendo con gran eficiencia ese objetivo original, o el recientemente (mayo de 2001) recomendado esquema de W3C, XSD que está mayormente enfocada a datos.

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:template match="articulo.1">
  <HTML>
    <HEAD>
      <META content="text/html;charset=ISO-8859-1"/>
      <TITLE>Revistas de Educación y pedagogía </TITLE>
    </HEAD>
    <BODY>
      <H1 align="CENTER">Revistas de Educacion</H1>
      <H2 align="CENTER"> <xsl:value-of select="relaciones"/> </H2>
      <H3 align="CENTER"> <xsl:value-of select="portada/titulo"/> <BR/><BR/>
        <xsl:value-of select="portada//nombre"/> <BR/>
        <xsl:value-of select="portada//cargo"/> <BR/>
        <xsl:value-of select="portada//institucion"/> <BR/>
        <xsl:for-each select="portada//email">
          <xsl:value-of select="."/> <BR/>
        </xsl:for-each>
    </H3>
     <H4>
       <xsl:for-each select="portada//resumen/p">
         Resumen:
         <xsl:value-of select="."/>
       </xsl:for-each> <BR/> <BR/>
     </H4>
      <HR />
        Copyright 2003 <BR />
         Alma Beatriz Rivera Aguilera, Universidad Iberoamericana  <BR />
         <A HREF="mailto:alma.rivera@uia.mx"> alma.rivera@uia.mx</A>
    </BODY>
  </HTML>
        </xsl:template>
```

```
            <xsl:value-of select="."/>
        </xsl:template -->
</xsl:stylesheet>
```

Figura 2.8. Ejemplo de hoja de estilo XSL.

2.4.2 Esquemas: DTD y XSD

El registro del diseño y funcionalidad de un vocabulario XML se encuentra depositado en su esquema específico, ya sea en la forma de DTD o de XSD (también llamado XML Esquema). Cómo ya se indicó en el apartado 1.6 la diferencia entre DTD y XSD es en relación a la sintaxis de definición de las etiquetas que conforman el vocabulario, que para XSD es similar a un archivo XML y para el caso de DTD tiene sus propias particularidades, (Ahmed, K. et.al., 2001 y Maler, E. y El Andaloussi, J., 1996).

```
<!ELEMENT articulo.1 (relaciones, portada, texto)>
<!ELEMENT relaciones (revista, ejemplar)>
<!ELEMENT revista (#PCDATA)>
<!ELEMENT ejemplar (#PCDATA)>
<!ELEMENT portada (autor+, titulo, fuenteBib?, fechaRecibido?, fechaAceptado?, epigrafe?, resumen?)>
<!ELEMENT fuente (itemBib)>
<!ELEMENT autor (#PCDATA | nombre | cargo | institucion | email)*>
<!ELEMENT nombre (#PCDATA)>
<!ELEMENT cargo (#PCDATA)>
<!ELEMENT institucion (#PCDATA)>
<!ELEMENT email (#PCDATA)>
<!ELEMENT titulo (#PCDATA)>
<!ELEMENT fechaRecibido (#PCDATA)>
<!ELEMENT fechaAceptado (#PCDATA)>
<!ELEMENT epigrafe (#PCDATA )>
<!ELEMENT resumen (#PCDATA )>
```

Figura. 2.9. Ejemplo de DTD.

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
   <xsd:element name="articulo" type="TipoArticulo" />
      <xsd:complexType name="TipoArticulo"/>
        <xsd:sequence>
            <xsd:element name="relaciones" type="TipoRelaciones"
                  minOccurs="1" maxOccurs="1"/>
               <xsd:element name="portada" type="TipoPortada"
                  minOccurs="1" maxOccurs="1"/>
               <xsd:element name="texto" type="TipoTexto"
                  minOccurs="1" maxOccurs="1"/>
         </xsd:sequence>

 <xsd:complexType nombre="TipoPortada"/>
   <xsd:all>
     <xsd:element name="autor" type="TipoPersona"
                  minOccurs="0" maxOccurs="unbounded"/>
     <xsd:element name="titulo" type="xsd:string"
                  minOccurs="1" maxOccurs="1"/>
   </xsd:all>

<xsd:complexType nombre="TipoPersona"/>
   <xsd:sequence>
     <xsd:element name="nombre" type="xsd:string"
                  minOccurs="1" maxOccurs="1"/>
     <xsd:element name="cargo" type="xsd:string"
                  minOccurs="0" maxOccurs="1"/>
     <xsd:element name="insitucion" type="xsd:string"
                  minOccurs="0" maxOccurs="1"/>
     <xsd:element name="email" type="xsd:string"
                  minOccurs="0" maxOccurs="1"/>
   </xsd:sequence>

…… Defínción del resto de tipos  …….

</xsd:schema>
```

Figura 2.10. Ejemplo de XML Schema también conocido como XSD.

Algunas comunidades se han puesto de acuerdo en el uso de etiquetas y tienen vocabularios comunes tales como el CML (Chemical Markup Language). Las etiquetas que cada quien genere se documentan en un archivo DTD de forma que siguiendo el ejemplo de CML existe un archivo que recoge las reglas del etiquetado llamado cml.dtd de acuerdo con el cual se marcan los archivos XML correspondientes.

En el caso de esta investigación el lenguaje de marcado XML es el elegido para llevar a cabo la definición del vocabulario y los marcados de los artículos ya que es un estándar cada día más utilizado tanto en la industria como en la academia.

2.4.3 Estándares e Interoperabilidad

Como se ha dicho en apartados anteriores las bibliotecas digitales se componen de colecciones de archivos u objetos digitales y se necesita recuperar información inmersa en ellos; esto se hace a través de los metadatos y las estructuración de los objetos. XML es un marco flexible para describir estructuras documentales y metadatos, permite hacer ligas sencillas y sofisticadas y está apoyado por un conjunto de estándares como:

- **RDF** (Resource Description Framework) para describir el contenido de los documentos y que se verá con detalle en el siguiente apartado
- **XQuery** que es un marco poderoso para enviar solicitudes de información y recibir los resultados en forma de listas o de documentos XML.
- **Open eBooks** como un repositorio de contenidos y secuencias de lectura.

XML por todo lo ya descrito permite la interoperabilidad entre los sistemas, siempre y cuando se respeten los estándares propuestos. Diseños conceptuales que se han ya implementado en diferentes lenguajes para aprovechar la interoperabilidad son el protocolo Z39.59 y el del Open Archive Initiative, este último tiene muy en cuenta los repositorios de metadatos que utilizan el XML. (Witten, I.H. y Bainbridge, D., 2003, cap.8)

2.4.4 Espacio de nombres[32]

"La premisa principal de XML es la creación de vocabularios de marcado formados por elementos personalizados (etiquetas), cuando esto se da en un entorno cerrado, la asignación de nombres a los elementos no es muy importante, ya que se pueden crear nombres de elementos únicos. No obstante, cuando se mezclan muchos vocabularios XML personalizados en la Web, la asignación de nombres únicos resulta una tarea difícil (si no imposible) sin contar con algún esquema de asignación de nombres adicional. Afortunadamente, existe ese esquema que involucra a los espacios de nombres" (Morrison, M. et.al., 2000, p. 110).

El concepto de espacio de nombres elimina las ambigüedades en relación a los nombres de etiquetas asociando una URI (Uniform Resource Identifier)[33] para cada aplicación XML y añadiendo un prefijo a cada elemento para indicar a cual aplicación pertenece (Harold, E.R., 2001, 331 p.)

Algunos ejemplos de espacios de nombres son:

```
xmlns:rdf   = "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
          rdf:Description, rdf:value

xmlns:dc    = "http://purl.org/dc/elements/1.1/"
          dc:title, dc:description, dc:creator, dc:type

xmlns:dcq  ="http://purl.org/dc/quaifiers/1.0"
          dcq :scheme, dcq :hasPart

xmlns:xsl   = http://www.w3.org/1999/XSL/Transform
          xsl :stylesheet, xsl :template
```

Figura 2.11 Espacios de nombre

[32] En XML el concepto "namespace" difiere al de programación, que se refiere a una biblioteca de clases necesaria cuando se crean varias clases con el mismo nombre pero con propósitos diferentes. La agrupación lógica de clases relacionas en un espacio de nombre tiene como propósito evitar colisiones de clases con el mismo nombre.

[33] URI engloba a los URL y URN. Los primeros son los más comunes direcciones de recursos en Internet y son dependientes de la ubicación, las URN (Universal Resource name) garantizan la singularidad de una dirección y tiene un nombre único independiente de la ubicación física del recurso.

La primera línea del ejemplo corresponde al espacio de nombre para etiquetas RDF, el segundo para los elementos Dublin Core, el tercero para los elementos calificados del mismo vocabulario y finalmente el espacio de nombre para las hojas de estilo XSL que al escribirse como archivos XML pueden hacer uso del espacio de nombres como se ve más ampliamente en la figura 2.8.

Cuando un vocabulario se define como novedoso y único se dice que puede generar un espacio de nombre (namespace); cuando un vocabulario utiliza etiquetas ya definidas por varios espacios de nombres se dice que genera un perfil de aplicación (application profile).

Para efectos de esta investigación el concepto de espacio de nombre es muy importante pues permitiría registrar el vocabulario a proponer, pero este proceso requeriría una consenso de la comunidad editorial, responsable de la publicación de revistas académicas, para llevarse a cabo con éxito. En el proceso de diseño se considerará si el vocabulario será un potencial espacio de nombre o un perfil de aplicación.

En este apartado se han visto las características principales de XML, sus elementos, atributos y entidades; como pueden desplegarse los archivos .xml a través de las hojas de estilo, como los estándares asociados al XML facilitan la interoperabilidad y finalmente la globalización de las aplicaciones a través del uso de espacio de nombre.

Comprender el espacio de nombre facilita el acercamiento al RDF uno de los estándares más pertinentes para el tema de la recuperación de contenidos inmersos en objetos digitales a través del uso de metadatos.

2.4.5 RDF como esquema de lenguaje de metadatos descriptivos.

RDF es el acrónimo de Resource Description Framework y se trata de una recomendación de W3C que establece un Marco para la Descripción de Recursos en la web a través de metadatos[34]. Se trata de un vocabulario XML que proporciona un modelo para registrar información descriptiva de recursos web que facilita la recuperación de los mismos, la información concreta irá vaciada en algún vocabulario específico de metadatos como por ejemplo el Núcleo de Dublín (Dublín Core). El concepto de RDF asume toda la web como el gran repositorio de colecciones digitales y a través de su definición nos facilita la asignación de metadatos a través de etiquetas XML normalizadas, que no estandarizadas; es decir con una estructura prediseñada, pero con nombres de etiquetas y especificaciones de contenido que el usuario o comunidad de usuarios pueden definir. (Lassila, O.,1999; Ahmed, et. al., 2001, cap. 4, 5 y 6; Harold, E.R, 2001, cap. 21).

Un ejemplo simpático, pero muy claro es el de Harold, E.R. (2001, p.707) quien define RDF como una aplicación de XML para codificar metadatos particularmente hecha para describir sitios y páginas web de forma que un buscador pueda hacer su trabajo lo mejor posible y no confundir a Homero el padre de la literatura occidental con el papá de Bart Simpson

[34] Las especificaciones técnicas pueden encontrarse en http://www.w3.org/1999/02/22-rdf-syntax-ns#

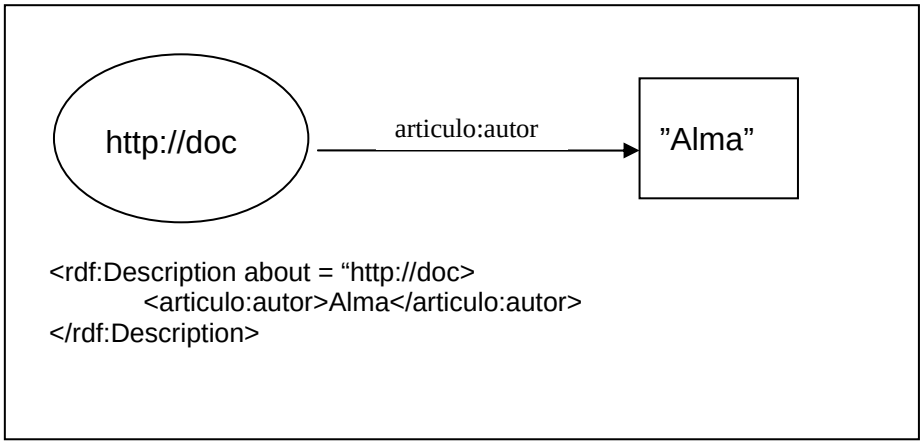


Figura 2.12 Modelo RDF. (tomado en parte de Miller, E. y Hillman, D., 2002, p.60)

RDF es el estándar de diseño que permite concretar mucho de lo que XML promete (Ahmed, et.al. 2000, p.97) y está compuesto por: (Miller, E. y Hillman, D., 2002, 60-61)

- Un **modelo de datos** que se conforma de dos recursos y su correspondiente relación identificada de forma única. La implementación de este modelo en XML permite la transmisión de datos.
- Un **esquema** que predefine aspectos propios de etiquetas que se pueden usar con fines de intercambio. El esquema permite a comunidades específicas, por ejemplo a las bibliotecas con el ya mencionado Dublín Core, crear vocabularios de metadatos que permiten interoperabilidad entre repositorios digitales.

En la figura 2.12 puede verse como el marcado RDF permite dentro de su contenido de elemento “Description” con el atributo “about” el uso de cualquier esquema de metadatos cuyo espacio de nombre haya sido registrado al inicio del archivo. Si el esquema referido a tarjetas de presentación virtuales

(xmlns:vcard="http://www.imc.org/pdi/Vcard/") hubiese sido integrado en ese ejemplo el archivo pudiese haber quedado:

```
<rdf:Description about = "http://doc>
        <dc:creator>
                <vcard:fn>Alma Rivera</vcard:fn>
                <vcard:org>Universidad Iberoamericana</vcard:org>
                <vcard:email>alma.rivera@uia.mx<vcard:email>
                <vcard:tel-work>59504000<vcard:tel-work>
        </dc:creator>
</rdf:Description>
```

Figura 2.13. Marcado RDF anidando el vocabulario Dublín Core y Virtual Card.

Un elemento importante señalado por Miller y Hillman es que sólo las comunidades pueden definir la semántica, es decir las etiquetas mismas y los alcances de una aplicación basada en RDF, en el caso de esta investigación para que el vocabulario propuesto fuera valioso en términos de interoperabilidad de sistemas se necesitaría un acuerdo entre los creadores de objetos digitales artículos de revista para que se utilizara las mismas o equivalentes etiquetas de marcado.

Marko, L. (2002, p. 57) nos indica que si HTML permite intercambiar documentos y XML definir etiquetas propias, RDF es lo que permite intercambiar información descriptiva o metadatos. Siguiendo con su analogía el autor refiere que a nivel de semántica existen aplicaciones específicas de RDF, una vez más con el ejemplo Dublin Core; la estructura está indicada por RDF y la sintaxis por XML. (Idem, p.59)

```
<?xml versión= “1.0”?>
<rdf:RDF>
  xmlns:rdf   = “http://www.w3.org/1999/02/22-rdf-syntax-ns#”
  xmlns:dc    = “http://purl.org/dc/elements/1.1/”
  xmlns:dcq   = “http://purl.org/dc/qualifiers/1.0/”
  <rdf:Description about=”urn:x-rslpcd:967715792-47835”>
    <dc:title>Colección Morrison de libros chinos</dc:title>
    <dc:description>
        Esta colección comprende libros chinos adquiridos por el Dr. Robert Morrison
     </dc:description>
     <dc:subject>
  </rdf:Description>
</rdf:RDF>
```

Figura 2.14 Ejemplo de marcado con espacio de nombre integrando tres diferentes vocabularios: el RDF, el Dublín Core y el Dublín Core Calificado.

Tenemos así que una herramienta que debe considerarse para establecer un vocabulario de metadatos descriptivos es el RDF. Algunos vocabularios existentes lo aplican como veremos en el apartado 2.6. Dado que esta investigación se enfoca en los metadatos estructurales más que un RDF se necesitaría un modelo del tipo “Resource Structural Framework” para diseñar el vocabulario en su parte estructural, el cual no existe.

Una propuesta interesante en relación a los metadatos descriptivos es la de los Mapas Temáticos (Ahmed, et.al. 2001, caps. 7 y 11), en la que no profundizaremos pues no refiere a aspectos estructurales de los documentos.

2.4.6 Vocabularios XML

Se ha revisado hasta ahora como las bibliotecas digitales se conforman de colecciones digitales, en las cuales el texto ocupa un lugar privilegiado, aunque no solitario. Para poder acceder a los contenidos de los documentos electrónicos se hace uso tanto de metadatos descriptivos como de búsquedas sobre el contenido completo, ya sea a través de bases de datos o buscadores en texto completo. Los

lenguajes de marcado facilitan, entre otras cosas, la recuperación tanto de los metadatos inmersos o asociados al texto como de los contenidos mismos y poseen los mecanismos para que estas aplicaciones de búsqueda corran sin confusión en toda la web. Algunos estándares como el RDF facilitan el intercambio de metadatos descriptivos.

En este apartado se hará una revisión de aplicaciones específicas de metadatos descriptivos y estructurales que de alguna manera pueden servir de modelo para la aplicación planteada por esta investigación.

Witten, I.H. y Bainbridge D. (2003 p.253-261) reportan como los estándares de metadatos más conocidos y utilizados en respositorios de documentos bibliográficos : el formato MARC, el Dublín Core, BibTeX y Refer. Las colecciones completas de documentos también puede ser descritas a través de metadatos y en esta línea se puede mecionar el EAD (Encoded Archival Innitiative) que es un esquema que describe a nivel colección los materiales de los archivos Kiesling, K. (2002).

Todos estos estándares representas esquemas de datos que pueden ser implementados a través de la tecnología de bases de datos, así como de lenguajes de marcado como el XML. En este apartado, y en esta investigación, nos enfocaremos a esquemas que son estándares oficiales o de facto y que se implementan comúnmente utilizando el XML.

Existen actualmente vocabularios que enmarcan contenidos descriptivos para objetos digitales, a los que hemos llamado durante todo este capítulo metadatos a secas o metadatos descriptivos, los cuales tienen varios usos en el web:

organizar, buscar, filtrar y personalizar sitios. Para efectos de este trabajo el enfoque es en el uso de lenguajes de marcado para generar un vocabulario estructural para potenciar la mejora en los resultados de la búsqueda de información con el fin de obtener mejores niveles de precisión. De acuerdo a Miller, Miller, E. y Hillmann, D. (2002) en el ámbito del uso de metadatos para la recuperación de información ha habido avance más bien la parte de los descriptivos que en relación a los estructurales.

2.4.6.1 Vocabularios descriptivos

- Para la descripción de archivos u objetos digitales, el vocabulario más utilizado es **Dublín Core**[35] que podríamos definir como "una colección de elementos diseñados para ayudar a los investigadores a encontrar recursos electrónicos de forma similar a la de un sistema de biblioteca" (Harold, E.R., 2001, p. 712). Las etiquetas de este vocabulario incluyen: titulo, creador, tema, descripción, editor, fecha, tipo, formato, identificador, fuente, idioma y relación. El conjunto de etiquetas DC fue aceptado como estándar de la ANSI en octubre de 2001. (Ver apéndice donde se describe el vocabulario).
- **GILS** (Global Información Locator Service)[36]. Esta tecnología es descrita en su sitio web (GILS : about a powerful..., s.f.) como un estándar abierto para la búsqueda de información descriptiva básica de recursos web. Es un esquema muy sencillo que incluye elementos, estructura de índice y los operadores a utilizar para las búsquedas.

[35] http://uk.dublincore.org/schemas/xmls/, utilzado como metadatos en muchos bibliotecas digitales
[36] htto://www.gils.net/about.htm.

2.4.6.2 Vocabularios descriptivos/estructurales

Al igual que las etiquetas de metadatos registradas a través de lenguaje de marcado y con un enfoque bibliográfico, como sería el caso de Dublín Core, los contenidos pueden marcarse a nivel de estructura de los documentos utilizando etiquetas inmersas en el contenido. Ejemplo de este tipo de vocabulario son:

- **TEI** (Text Encoding Innitiative)[37] combina un conjunto de etiquetas que incluye tanto las descriptivas como las de tipo estructural. Se desarrolló en SGML desde 1987 hasta la publicación de las Normas en 1994 y está diseñada para textos académicos en el área de humanidades. Es un esfuerzo internacional e interinstitucional de la Association of computers and the Humanities, la Association for Computer Linguistics y la Association for Literary and Linguistic Computing. Existe una versión llamada TEI-Lite con menos etiquetas que la general y adaptaciones en XML. (Muller, M, Introducción, párr. 4 y Burnard, L. y Sperberg-McQueen, C.M. 1995). Marko, L. (2002, p. 53) señala que en la medida que el análisis de texto se mueva a la web será mejor tener estándares para llevarlo a cabo, en este caso el DTD del TEI es una buena aproximación tomada por varias instituciones. Esta autora se interesa fundamentalmente por el llamado TEI Header, es decir la porción del esquema referido a datos descriptivos. Se considera que es bueno evaluar la adopción de la propuesta de manera integral para lograr el beneficio de los datos estructurales. El TEI en esquema complejo de 450 elementos que se ha simplificado en versiones

[37] http://www.tei-c.org.uk/, utilizado en análisis de textos

llamadas TEI Lite con 150 elementos y TeiXBaby de 60 elementos (Muller, The XML versión of TEI, párr. 2). El TEI Lite es el más recomendado por los autores y (Marko 2002, p.54 y Muller). El éxito futuro de este estándar depende de la cooperación de los generadores de documentos en el uso del marcado para poder trabajar en análisis lingüísticos con variedad de fuentes en Internet. Fietzer, W., (2002, p. 103) menciona como el marcado estructural no da sólo mejoras en la búsqueda sino posibilidades de análisis literario. Menciona el TEI como una oportunidad de colaboración entre los sistemas de recuperación de información con la academia.

- **MOA2** (Making of America II)[38] /**METS** (Metadata Encoding and Transmission Standard)[39]. MOA es un proyecto de la Digital Library Federation en la universidad de Berkeley, específicamente en la biblioteca Bancroft y propone un DTD de XML que permite marcado de metadatos descriptivos, administrativos y estructurales. Está enfocado a documentos propios de archivos. (Beaubien, R., 2001). METS retoma la propuesta del MOA2 y provee el formato de documentos XML para codificar los metadatos necesarios tanto para la administración de objetos digitales dentro de un repositorio como para el intercambio de dichos objetos entre diferentes repositorios, actualmente se encuentra administrado por la Biblioteca del Congreso de los Estados Unidos
- **SciELO** (Scientific Electronic Library Online)[40]. Este esquema es una propuesta brasileña generado en BIREME (Centro Latinoamericana y del

[38] http://sunsite.berkeley.edu/moa2/
[39] http://www.loc.gov/standards/mets/METSOverview.v2.html

Caribe de Información en Ciencias de la Salud) parte de la OPS/OMS (Organización Panamericana de la Salud/Organización Mundial de la Salud). Está enfocado a artículos de revista y tiene gran elaboración a nivel descriptivo, la estructura está volcada en un solo elemento por lo que las búsquedas a nivel de contenido no refieren a ningún marcado. (Marcondes, C.H. y Sayao, L.F., 2003 y SciELO Metodología, 2001)

- **JAI/JP**. Journal Archiving Interchange DTD y Journal Publishing DTD. Vocabularios propuestos por el Centro Nacional de Información sobre Biotecnología (NCBI) de la Biblioteca Nacional de Medicina (NLM) de los Estados Unidos. El DTD del JAI describe artículos de revista, reseñas de productos y de libros con el proposito de almacenarlos e intercambiarlos en formato sitial. El esquema JP es un subconjunto de JAI y se propone para revistas que no tengan un modelo XML/SGML establecido y que aportan sus contenidos en PubMed Central[41]

Para los vocabularios de artículos las comunidades de usuarios han hecho propuestas para el diseño e implementen los esquemas de metadatos correspondientes, como en el caso del TEI, el JAI o Scielo. Sin embargo, de forma más interna empresas comerciales como EBSCO han desarrollado esquemas propios, y aunque no poseen todo su material de revistas electrónicas marcado han desarrollado esquemas de marcado que utilizan en su mayoría para despliegue más que para recuperación[42] y que de alguna manera poseen mayor impacto pues este tipo de empresas son las que están generando mayores

[40] http://www.scielo.org/dtd/
[41] Archivo digital de la NLM con artículos de revistas sobre ciencias de la vida

repositorios de artículos y llegando a más cantidad de usuarios a través de sus paquetes de colecciones de revista como Academic Search Premier[43] o Academic Literature[44]

Una vez revisados los vocabularios tanto descriptivos como los que combinan descripción y estructura, debe señalarse que se pueden combinar estándares de metadatos, por ejemplo un marcado como Dublín Core o RDF pueden estar inmersas en un mismo esquema (ver Figuras 2.13 y 2.14). Un punto importante a señalar es la variedad de propuestas de marcado y los esfuerzos de integración a través búsquedas federadas. No queda claro aun si lo más eficiente sería establecer estándares de marcado estructural como los hay ya de descriptivo o desarrollar herramientas que aprovechen cualquier tipo de marcado. Por la experiencia generalizada en la web lo ideal son los estándares, pero para llegar a ellos siempre se toma algún tiempo y mucha voluntad de los protagonistas.

2.5 Búsqueda e indizado

Los organismos vivos estamos todo el tiempo buscando información sobre nuestro ambiente y mientras más complejos son los organismos, más complejas son las estructuras cognitivas que facilitan la búsqueda (Belew, R.K., 2000, p. 2).

De acuerdo a este mismo autor al buscar utilizamos uno de los ambientes más complejos: el lenguaje; y todo proceso de búsqueda requiere 3 fases (Idem, p. 5-8):

1. Preguntar

---

[42] Pesch, O. (2003, 22 de enero). Comunicación personal por e-mail.
[43] http://www.ebsco.com
[44] http://www.proquest.com

2. Construir una respuesta
3. Evaluar la respuesta

El tema de la búsqueda y recuperación de información ha sido fundamental para las ciencias de la computación, las cuales incluyen al concepto de recuperación de información (RI equivalente a IR Information Retrieval) y al de bases de datos (BD equivalente a DB Data Bases) como dos sub disciplinas consolidadas dentro de su corpus teórico.

La información de tipo textual que en esta investigación nos ocupa posee elementos de contenido recuperables integrados en el texto mismo y la posibilidad de identificar y vaciar como elementos buscables los llamados metadatos descriptivos. Estos últimos pueden estar en una base de datos, estar integrados al documento mismo a través de un marcado inicial o encontrarse en otro documento que sólo contiene metadatos y hace referencia a un URL donde se encuentra el documento principal.

Como se indicó antes el proceso de búsqueda y recuperación de información es complejo y al abordarse en los textos puede servirse de herramientas automatizadas desde las más simples hasta las más sofisticadas, desde sencillos archivos invertidos sobre textos planos hasta agentes y arañas en la web, pasando por las consolidadas tecnologías de bases de datos.

Con respecto al contenido mismo de un material en texto completo, este puede ser recuperable a través los ya mencionados índices, los cuales asocian palabras o frases a un documento y su generación puede ser manual (se han hecho índices desde siglos antes de la existencia de las computadoras) o automáticos.

Las formas más comunes de llevar a cabo los índices son, de acuerdo con Witten, I.H, Moffat, A. y Bell, T.C., (1999, cap. 3)[45]:

1. Archivos Invertidos
2. Archivos “firmados”
3. Bitmaps

Belew propone el uso de elementos de inteligencia artificial como el área de aprendizaje de máquina como una herramienta de apoyo a la búsqueda más adecuada. Este autor no menciona las posibilidades del lenguaje de marcado para mejorar la búsqueda, indica que lo más conveniente es eliminar el marcado (HTML o XML) de los documentos para facilitar el proceso de los indizadores (Idem, p. 41). En esta investigación se pretende aprovechar el marcado como una herramienta para mejorar los resultados de una búsqueda.

En los siguientes apartados se revisará el papel de la tecnología de bases de datos y los modelos de indizado para documentos XML para la recuperación de información.

2.5.1 Bases de datos

De acuerdo a Elmasri, S. y Navathe, S.B. (2000, p. 4-5) una base de datos es una colección de datos relacionados. Esta definición es de tipo general y podría generar confusiones e incluso aplicarse a un párrafo compuesto de palabras. De forma más restringida una base de datos tiene las siguientes propiedades:

1. Representa algún aspecto del mundo real.

[45] El tema de los índices y buscadores es enorme y no se tratará en detalle en esta investigación, se aprovecharán herramientas disponibles de software gratuito para las búsquedas.

2. Es una colección lógicamente coherente de datos con un significado inherente. Colecciones de datos aleatorios no se consideran base de datos.
3. Esta diseñada, construida y poblada de datos con un propósito específico. Existe un grupo de potenciales usuarios y algunas aplicaciones que se supone interesarán a dichos usuarios.

Por ejemplo, un catálogo de libros es una base de datos que incluye datos (llamados campos) sobre autores, títulos, editores, año de publicación, etc. Dichos datos se relacionan al representar un solo ítem de información llamado registro. Cada libro sería equivalente a un registro y sus campos asociados serían el autor del libro en particular, el título, editorial, año de publicación. etc.

Las bases de datos han jugado un papel muy importante como repositorios de información que facilitan el almacenamiento y recuperación de datos. En el ámbito de las bibliotecas han cumplido desde los años 50´s un papel fundamental en la automatización de catálogos de colecciones locales y de referencias bibliográficas de materiales más allá de los muros de dichas instituciones. Las bibliotecas digitales han aprovechado esta tecnología para almacenar los catálogos de los objetos digitales, al almacenar en un sistema administrador de bases de datos los llamados metadatos como campos. Cabe aclarar que los metadatos no necesariamente deben almacenarse en una base de datos.

Otra vía de recuperación de información es la utilización de buscadores sobre textos completos. Para hacer eficiente la búsqueda sobre dichos textos se utiliza el concepto de metadatos descriptivos de los materiales, los cuales se pueden encontrar incrustados o asociados a los objetos digitales de tipo textual y que pueden ser utilizados por buscadores especializados en este tipo de metadatos

para recuperar información. Estos buscadores especializados hacen uso de las marcas inmersas en el texto para identificar cuales son las cadenas de caracteres que representan por ejemplo una autor, título o tema. Las marcas se colocan atendiendo a las reglas de los llamados lenguajes de marcado que se describen en el siguiente apartado.

2.5.2 Buscadores en texto completo y para XML

Como se mencionó al inicio del apartado II.5 los formatos comunes de generación de índices para textos completos son los archivos invertidos, los de firma y los llamados bitmap. Estos conceptos generales pueden aplicarse tanto para la asociación de palabras a registros, tuplas u objetos en una bases de datos como a un documento en texto completo. En el caso de documentos de acuerdo al diseño del proceso de indizado la asociación puede variar de acuerdo a la granularidad deseada, es decir, el elemento del índice puede asociarse a un documento, a una parte del documento, a un párrafo, o a una frase.

Luk, R.W.P., et.al. (2002, p. 416-422) hacen una revisión exhaustiva de las diversas técnicas de indizado y búsqueda para documentos XML[46]. Estos autores proponen una clasificación de las técnicas de generación de índices basada en las características de marcado de los documentos:

1. Indizado de Archivos Planos. Utiliza herramientas convencionales de RI que desafortunadamente pierden la riqueza del marcado, se han utilizado

[46] El volumen 53 de abril de 2002 del Journal de la American Society for Information Science está dedicado al XML y cuenta en especial con artículos en relación a la recuperación de documentos en XML.

manipulaciones para aprovechar las etiquetas como elementos de índice y posibilitar combinaciones. (Idem, p.416)

2. Indizado Semi Estructurado:
    a. Basado en campos, los términos del índice se construyen combinando el nombre del campo con las palabras del contenido.
    b. Basado en segmentos, los documentos son divididos en regiones
    c. Basado en árboles, dado que la estructura de los documentos es jerárquica, a cada nodo se le asigna un identificador único.
3. Indizado Estructurado
    a. Combinación de técnicas de RI/BD, como el conocido indizado secuencial (indexed sequential file organization) ISAM y su generalización VSAM utilizando el modelo de B+tree para mantener los registros actualizados fácilmente. Con la combinación de XML y modelos de datos de BD los documentos pueden almacenarse en campos BLOB (Bynary Large Objects) y aprovechar los procesos de la base de datos. Uso del DOM y el SQL (Idem, p. 418–419).
    b. Basado en direcciones (“path”), utilizando bases de datos orientadas a objetos (OODB) se construyen índices basados en el diccionario de “paths” (Idem, p.421)
    c. Basado en la posición, ve el archivo como un objeto de dos dimensiones, en el cual se identifican regiones rectangulares a

través de las etiquetas. El índice se aplica al documento original y cualquier versión.

d. Multidimensional, índices independientes por elementos a través de B-Trees o R-Trees. Para el caso de cubos de datos utiliza utilerías OLAP (On Line Analytical Processing)

En esta investigación no se ahondará en detalle en estas tecnologías ya que el objetivo de investigación es proponer un lenguaje de marcado y no un buscador; pero cabe recordar que la hipótesis principal de esta investigación es que los documentos marcados generarán la posibilidad de una mejor recuperación y para poder explorar la validez de dicha hipótesis será necesario utilizar un buscador sobre textos completos que tenga la opción de indizar el marcado. Luk, R.W.P., et. al. (2000, p. 433-434) consideran que para que XML cumpla la promesa de resultados de búsqueda más precisos, búsquedas integradas de fuentes heterogéneas, búsquedas más poderosas utilizando especificaciones estructurales y de contenido e intercambio de datos en apoyo a búsquedas cooperativas debe investigarse más sobre diversos tópicos como la heterogeneidad de los datos, el ordenamiento de resultados (ranking), la evaluación de resultados, los modelos de recuperación, el indizado, la búsqueda y la administración de documentos. En esta investigación, como ya se ha dicho en diversas ocasiones, se hará referencia a una propuesta de homogenización de marcado y se revisará la evaluación de la precisión.

Luk, R.W.P. (Idem) no reportan el valor de los vocabularios genéricos y podríamos afirmar que no hay un modelo ideal de buscador para XML, lo cual hace

interesante que al combinar un diseño de marcado adecuado a la colección de documentos con los modelos de recuperación adecuados puede lograrse que el XML realmente cumpla con lo prometido de ser un estándar adecuado para la mejor recuperación.

Búsqueda federada

Si bien en este trabajo se hace énfasis en el uso del XML por su impacto en la mejora de la precisión de los resultados de una búsqueda, una de las grandes virtudes de este lenguaje de marcado es la facilidad con que repositorios independientes e incluso heterogéneos pueden integrarse a través de la recolección (harvesting) de sus metadatos descriptivos. Para que dicha integración se de es necesario que existan programas que llevan a cabo la función de intercambio de metadatos, en esta línea vale la pena mencionar el proyecto Open Archive Innitiative[47] el cual propone un diseño estándar cliente servidor mediante el cual se pueden integrar metadatos de diferentes repositorios que describen colecciones digitales (MARC, Dublín Core, etc.) en un solo servidor, de forma que un usuario puede buscar información y obtener resultados a partir de diferentes colecciones. Las consideraciones del proyecto OAI son fundamentalmente en relación a metadatos de encabezado o descriptivos. Sería conveniente considerar en el futuro si un diseño similar se puede desarrollar para aprovechar marcados estructurales en diferentes repositorios.

[47] http://www.osi.org

2.5.3 Evaluación de la búsqueda

A pesar de que las medidas más populares para establecer la evaluación de los resultados de una búsqueda (Recall and Precision) fueron desarrolladas en los 50´s y se les ha criticado su subjetividad (Lancaster, F.W. 1983, p.161) y limitada visión (2002, Fuhr, N. et. al., p.188, 2001), todavía son usadas comunmente para evaluar los sistemas de recuperación de información (Belew, R. K, 2000, pp. 122-124 y Govert, N. y Kazai, G. 2003, pp. 8-9), tanto por los profesionales de la ciencias de la computación como por los bibliotecólogos.

La definición de precisión y relevancias puede verse en el Apartado 1.6 y su metodología es analizada en detalle en el Apartado 3.6.

Fuhr, N. et. al. (Idem, p.188) propone para la evaluación de resultados de búsqueda analizar el tipo de tareas que pretende resolver un repositorio de información digital e identificar parámetros de medición del éxito de la tarea, como por ejemplo, tiempo de obtención de resultados y porcentaje del problema realmente resuelto. Luk, R.W.P et. al. (Idem, p.433) señala la falta de colecciones suficientes de datos marcados en XML para ser evaluados, lo anterior debido principalmente a la relativa juventud del XML, a la conversión que en la WWW se da al HTML de los documentos en XML, el costo del marcado y el consecuente bajo nivel de adopción de esta tecnología por algunas instituciones que generan documentos digitales.

Belew, R. K (2000, cap. 5) dedica un capítulo de su publicación “Finding out about …” a la evaluación de los resultados de búsqueda, considera inicialmente que la evaluación es vista como una actividad de las personas y propone que el concepto

de Relevance Feedback (RelFbk) pueda ser aprovechado por los sistemas a al combinar la relevancia asignada a un material por múltiples individuos generando un RelFbk consensuado con validez estadística. La relevancia, según este autor, es un dato discreto y puede ser registrado con la escala: no relevante, no contesta, posiblemente relevante, relevante y críticamente relevante.

En esta investigación se utilizará la medida de precisión utilizando la evaluación de un documento como relevante o no relevante por individuos experto en el contenido de los textos.

2.6 Publicación digital

En los apartados anteriores se han revisado conceptos de biblioteca digital, lenguajes de marcado, búsqueda y recuperación de información y finalmente en esta última parte del capítulo 2 se hará una breve revisión de la publicación digital, también conocida como publicación electrónica; en especial la correspondiente a las revistas. En los últimos años las publicaciones periódicas en formato electrónico han sufrido un crecimiento exponencial, que va desde unas decenas de títulos en 1996 a cerca de 13, 000 para el año 2003; sin embargo, persiste la polémica sobre la persistencia del formato impreso. A diferencia del libro electrónico que no ha cumplido las expectativas comerciales de los editores, la revista digital se mantiene muy fuerte tanto en la oferta editorial tradicional a través de suscripciones y paquetes de bases de datos, como en la de los autores y sociedades académicas que promueven el acceso libre a los contenidos.

Sally Morris (2002, párrafo 2.) comenta que la revista digital tiene 4 grandes ventajas sobre las versiones impresas, pero para cada una de las ventajas hace una reflexión que la condiciona:

1. Cobertura internacional, siempre y cuando tenga promoción
2. Velocidad de publicación, pero si se quieren mantener los niveles de calidad el tiempo editorial se reduce sólo en la parte correspondiente a impresión y distribución pues la evaluación de pares y revisiones tomarán el mismo tiempo que la versión impresa.
3. Capacidades adicionales (vínculos, animación, sonido, etc.) Reportes sobre de usuarios (Swan, A. y Brown, B., 2002 y Baldwin, C. y Pullinger, D., 2000, citados por Morris, S.) registran que lo más valorado son vínculos por lo que hay que evaluar si otras capacidades como la multimedia merecen el tiempo y gasto requerido.
4. Bajo costo, se ahorra en costos de impresión cuando se opta por eliminar la versión en papel: sin embargo, los costos de manejo de datos y administración generalmente aumentan.

Otros beneficios a considerar son la comodidad de acceder 24 horas, 7 días a la semana; y la facilidad en la búsqueda de los textos y metadatos. Un aspecto que más bien preocupa es la capacidad de archivar en el tiempo o preservar los ejemplares ya que los soportes físicos digitales todavía no han demostrado su longevidad y el acceso a las revistas digitales es a través de consultas en sitios web la cual puede ser bloqueada si no se renuevan las suscripciones o simplemente desaparecer si se trata de sitios gratuitos.

Los aspectos de mercadeo, control de accesos, preservación del formato digital, etc. todos ellos importantes en la publicación digital de revistas, no son del interés específico de este trabajo y pueden consultarse en la extensa bibliografía sobre el tema que instituciones como el INASP (Internacional Network for the Availability of Scientific Publications) tienen disponible.[48]

Para efectos de esta investigación es muy importante que el vocabulario XML que se proponga para el marcado de artículos de revista no sólo se enfoque a identificar los elementos de recuperación de contenidos, sino que tenga en cuenta que los materiales deberán publicarse en un formato legible y agradable.

Los vocabularios mencionados en el apartado II.4.6 (MOA/METS, TEI, JAI) consideran las necesidades de la publicación digital para facilitar las diversas presentaciones de los datos haciendo uso de CSS o XSL necesarias así como los elementos de metadatos y contenido que facilitarán la recuperación y análisis literario, en el caso del TEI, de los textos.

Como se dijo al inicio de este capítulo el contenido aquí descrito es el marco temático para desarrollar el vocabulario XML a aplicar en artículos de revista que impacte en un mejor nivel de precisión de los resultados de búsqueda, dicho marco se ha construido con la revisión de las propuestas que nos ofrecen las grandes áreas del cómputo, la bibliotecología y la publicación en subdisciplinas de bibliotecas digitales, búsqueda y recuperación de información, lenguajes de marcado, catalogación, bases de datos y finalmente, pero no menos importante, la publicación electrónica de revistas.

[48] http://www.inasp.info

3. METODOLOGÍA

En el capítulo anterior identificamos cuáles son las principales tendencias en el ámbito del almacenamiento y recuperación de información en archivos textuales de bibliotecas digitales.

En este tercer capítulo retomaremos la propuesta de la sección 1.4 en donde se dio un esbozo general de la metodología a utilizar. Se retomará el problema e hipótesis de investigación, el tipo de investigación y el diseño preexperimental, la metodología para generar vocabularios XML y la metodología de evaluación de resultados de búsqueda.

3.1 Problema, objetivos, hipótesis

Como se registró en la introducción el problema de investigación que aquí nos preocupa nace de la dificultad real de recuperar información relevante y pertinente para los usuarios en grandes repositorios de documentos en texto completo, y puede considerarse como:

> *Definir un vocabulario XML que describa la estructura de artículos de revistas de tipo académico en bibliotecas digitales y recoja los contenidos significativos inmersos en el texto, de forma tal que la precisión en la recuperación de los resultados sea más alta que si un buscador indiza la totalidad de las palabras del texto.*

El objetivo principal, es

> *Proponer y evaluar un vocabulario XML que describa y estructure adecuadamente los contenidos de artículos de revistas académicas en formato digital, dicha vocabulario facilitará la diversidad de*

representaciones de tales objetos digitales y la recuperación de información de los mismos.

Un objetivo auxiliar es seleccionar y probar una metodología adecuada para el desarrollo de dicho vocabulario.

La hipótesis principal de la investigación es:

Al marcar artículos de revista con el vocabulario XML adecuado los resultados de la búsqueda, que aproveche el indizado sobre ciertas partes marcadas del texto de los artículos, serán más precisos que los resultados de la misma búsqueda sobre el texto completo.

Antes de pasar al diseño de investigación utilizada se señala que los artículos de revista y expertos evaluadores serán del área temática de educación, debido al acceso que de este tipo de material se tuvo de diferentes instituciones y a la facilidad de solicitar el apoyo de expertos en educación presentes en la comunidad UIA.

3.2 Tipo de investigación

Este estudio utiliza como marco metodológico para resolver el problema planteado en el apartado anterior a los dos tipos de investigación expuestos por Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (1998, cap.4) [1]:

- **Investigación Descriptiva**, ya que reporta el desarrollo de un vocabulario XML para textos completos de artículos de revistas y la metodología de marcado de los artículos con dicho vocabulario.

[1] En palabras de estos autores un estudio descriptivo se refiere a “cómo es y se manifiesta un fenómeno y sus componentes” y un exploratorio tiene como objetivo familiarizarnos con un tópico poco estudiado y sirve para desarrollar métodos a utilizar en estudios más profundos (p.71).

- **Investigación Exploratoria,** correspondiente a la medición de los resultados del marcado XML aprovechados por un buscador sobre la precisión en la recuperación de la información en artículos de revista.

Para el diseño del vocabulario se revisaron varias opciones metodológicas y se optó por la propuesta de Maler, E. y El Andaloussi, J. (1996) para la generación de DTD; metodología que aunque pensada para SGML se adapta perfectamente a los documentos textuales de estructura flexible ya sea en SGML o XML. Los criterios que llevaron a esta selección y los elementos propios de la metodología se detallan en el apartado 3.3

Las razones por las que la medición de la precisión en los resultados de búsqueda es exploratoria en este caso son:

- No existen metodologías y resultados definitivos[2] referidos a este tipo de estudio.
- No se contó con el tiempo y el personal necesario para llevar a cabo un marcado masivo de documentos necesarios para obtener resultados explicativos.

Un resultado explicativo hubiese requerido:

- La generación de una muestra probabilística, en este caso, una colección de artículos de revistas de educación elegidos al azar de entre la totalidad de revistas mexicanas sobre el tema educativo. Así como determinar el

[2] Hay ciertamente iniciativas llevándose a cabo actualmente como *Innitative for the Evaluation of XML retrieval (INEX)*, http://www.is.informatik.uni-duisburg.de/projects/inex03 ; con una metodología establecida. Sin embargo, INEX se enfoca al aprovechamiento de marcado con el que se cuenta por los buscadores, y este trabajo se centra en el vocabulario más adecuado. Ambos esufuerzos podrían considerarse complementarios.

número exacto de dichos artículos a fin de determinar el tamaño de la muestra.

- La conversión a formato digital de los artículos seleccionados.
- Marcado de artículos con el vocabulario propuesto.
- Identificación de los investigadores educativos de todo el país y selección de una muestra probabilística de dichos expertos.
- Desarrollo de un protocolo de evaluación que no requiriera la presencia física del investigador durante la búsqueda en el repositorio de textos.

Dado que lo anterior era difícil de establecer con los recursos de tiempo y personal disponible se optó por un enfoque exploratorio no probabilístico, y se diseñó una exploración del tipo preexperimento (Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. 1998, cap.6) [3].

Habiendo establecido que este estudio tiene en parte tratamiento descriptivo y en otra parte exploratorio, se tiene un preexperimento que ayude a responder a la hipótesis planteada

3.3 Diseño experimental

El preexperimento consistirá en ejecutar una estrategia de búsqueda sobre todo el texto completo y sólo sobre ciertas partes marcadas. Así tenemos:

Variable independiente: marcas en los artículos (presentes o ausentes)

Variable dependiente: nivel de precisión (a través de la evaluación de la relevancia de los artículos)

[3] Los preexperimentos son estudios exploratorios y descriptivos, y se consideran una primera aproximación al experimento pero sin validez probatoria. Los experimentos pueden definirse como: "estudios en donde se manipulan deliberadamente una o más variables independientes (causas)

El grado de manipulación de la variable independiente será el mínimo y se refiere a la presencia o ausencia, en este caso tomar en cuenta o no, el marcado en el momento de indizado y búsqueda. De igual forma la respuesta de los evaluadores con respecto a la relevancia de los documentos recuperados a través de la búsqueda serán de grado mínimo ya que se tratará de calificar de relevante o no relevante cada artículo del conjunto de resultados.

Las muestras no probabilísticas a utilizar son:

- Una colección de documentos marcados conformada por 29 artículos de revista sobre educación provenientes de las revistas:
    - DIDAC de la Universidad Iberoamericana, un número, 8 artículos
    - La Academia del Instituto Politécnico Nacional, dos números, 10 artículos
    - Pedagogía de la Universidad Pedagógica Nacional, dos números, 11 artículos.
- Un grupo de 8 expertos en educación de la comunidad de la UIA, como evaluadores.
- Cada evaluador propondrá 3 temas a buscar, se llevará a cabo la correspondiente estrategia de búsqueda y se tendrán 24 conjuntos de resultados a evaluar.

Cada tema filtrados por un buscador en la opción de texto completo, sobre el que se aplicarán estrategias de búsqueda asociadas a los temas propuestos y

---

para analizar las consecuencias sobre una o más variables dependientes (efectos) de la manipulación en una situación controlado por el investigador" (p.107)

establecidas de común acuerdo entre el investigador y los evaluadores[4]. Las estrategias de búsqueda utilizando palabras conectadas con operadores booleanos AND y OR. La presencia del investigador es importante para resolver cualquier duda o problema técnico de conexión o despliegue[5].

Sobre el conjunto de artículos recuperados el experto hará una valoración de relevancia uno a uno con respecto al tema buscado, indicando si es relevante o no, el investigador tomará nota del nombre del archivo y del juicio del experto.

Finalmente se identificarán en cada sets generados los artículos que contengan las palabras establecidas en la estrategia en ciertas etiquetas del texto y se retomará la valoración que el individuo dio a los artículos, para calcular los ratios de precisión, en ambos casos: búsqueda de texto completo y búsqueda sólo en etiquetas.

[4] Se puede dejar que el evaluador diseñe la estrategia de búsqueda por si solo, pero algunos de ellos no están familiarizados con uso de operadores y paréntesis.

[5] Este caso se dio pues algunos de los investigadores poseía Explorer 5 o menor, lo cual no permitió en todo los casos ver el despliegue generado por las hojas de estilo XSL.

Forma de Recolección de Datos de Medición de Precisión

Nombre del investigador: _____

Email: _____ Ext.: _____

Breve referencia de su formación académica y experiencia en el área de la educación: _____

Tema 1: _____

_____

Estrategia 1: _____

Artículos recuperados: _____ Artículos útiles: _____

Tema 2: _____

_____

Estrategia 2: _____

Artículos recuperados: _____ Artículos útiles: _____

Tema 3: _____

_____

Estrategia 3: _____

Artículos recuperados: _____ Artículos útiles: _____

Figura 3.1. Forma de recolección de datos para la medición de precisión.

La recopilación de los datos necesarios para el preexperimento se hará mediante visitas a los evaluadores durante las cuáles se llevarán a cabo los procesos de búsqueda y evaluación necesarios para el llenado de la forma

Pruebas con mayor cantidad de archivos serían alcanzables eventualmente pues los desarrollos técnicos ya están elaborados, y el marcado es en sí un ejercicio simple aunque minucioso y que debe supervisarse para asegurar la calidad.

Los resultados no podrán generalizarse pues el preexperimento no está llevado a cabo sobre muestras significativas ni de documentos, ni de evaluadores. El

análisis de los datos recolectados será útil como un primer acercamiento, que abre un camino al problema aquí planteado, pero serán necesarias estudios más completos en términos de colecciones y evaluadores; sin embargo se considera que el establecimiento del vocabulario, el buscador y la metodología de medición de precisión es un firme avance.

3.4 Metodología para la generación de un vocabulario XML

La revisión de la bibliografía referida a diseño de vocabularios XML llevó a la conclusión de que hay 3 posibles esquemas para el diseño: metodologías propias del lenguajes de marcado, diseño orientado a objetos y diseño de bases de datos. En el presente apartado se revisarán los 3 enfoques y se seleccionará el utilizado para esta investigación.

3.4.1 Metodología SGML/XML

Diversos autores (Maler, E. y El Andaloussi, J. (1996); Jeliffe, R. (1998).; Harold, E.R. (2001), Morrison, M. et. al. (2000) proponen cómo se debe generar un esquema ya sea del tipo DTD o del tipo XML Esquema. Harold lleva a cabo una propuesta práctica y sencilla de desarrollo de nuevas aplicaciones (pp. 995-1024). Morrison se enfocan a la sintaxis de la escritura del DTD o del XML Esquema y otros como Maler y El Andalousi y Jeliffe hacen una propuesta a nivel de ingeniería de software antes de entrar en los detalles de la implementación. Todos los autores introducen inicialmente los conceptos de elemento, atributo, entidad y comentario; dan ejemplos de archivos marcados y DTD´s y/o XML schemas y como se integran los comentarios (ver apartado II.4.1)

Harold, E.R. (2001) dedica todo el capítulo 28 al desarrollo de una nueva aplicación XML y considera que hay 3 pasos básicos en este proceso:

1. Listar los elementos
2. Identificar los elementos fundamentales
3. Relacionar los elementos unos con otros

Este autor menciona (Idem, p. 997), sin ahondar en detalles, que hay una gran similitud entre estos tres pasos y las técnicas que se utilizan para identificar requerimientos de los usuarios en programación orientada a objetos y diseño de base de datos.

Maler, E. y El Andaloussi, J. (1996) proponen una metodología muy completa que se conforma de 6 fases de diseño e implementación (p. 30-31)[6] y que se complementa en el texto con indicaciones para la administración del proyecto.

Los seis pasos de estas autoras se resumen en:

1. Articular los objetivos del proyecto
2. Analizar las necesidades que pretenden satisfacer los datos contenidos en los documentos
    a. Identificar y definir los componentes de información básica
    b. Clasificar los componentes en grupos lógicos
    c. Validar el análisis contra otros modelos existentes
3. Diseñar los requerimientos del esquema basándose en los objetivos. Recabar información sobre requerimientos con expertos, desarrolladores de las aplicaciones que procesarán los documentos y usuarios finales.

[6] En realidad gran parte del libro de Maler y El Andaloussi se trata del desarrollo de la metodología de los 6 pasos que en estas dos páginas citadas se encuentran resumidos.

a. Seleccionar los componentes semántico
b. Construir la jerarquía del documento y la metainformación.
c. Construir el modelo de elementos y atributos a nivel medio del documento: unidades de información
d. Construir el modelo de elementos y atributos para jerarquía de bajo nivel del documento.
e. Proponer variedad de elementos de donde un autor pueda seleccionar los necesarios.
f. Hacer conexiones entre el modelo y un modelo del mundo externo.
g. Validar que el modelo esté completo y que está tomando en cuenta otros modelos similares ya desarrollados.

4. Completar el diseño del esquema e implementarlo. En este punto decidir si tendrá un enfoque modular o se crearán varios esquemas que cubran variedad de documentos.
5. Validar el resultado y verificar que se están alcanzando los objetivos
6. Documentar el esquema y capacitar a la gente que lo utilizará

El trabajo de Maler, E. y El Andaloussi, J. detalla cada paso y está lleno de valiosos consejos para el análisis estructural de los documentos, sin embargo, no hace una especial referencia a otros modelos de desarrollo de software en los que seguramente se basa.

Jeliffe, R. (1998, cap.3) por el contrario hace una revisión de una gran variedad de métodos para el desarrollo de esquemas[7], los cuales organiza en 4 grupos:

[7] Jelliffe usa el término DTD para referirse a los esquemas de SGML, aquí se utiliza el término esquema, el cual incluye como resultado de la metodología tanto un DTD como un XSD.

1. **Generación del esquema a partir de componentes o arquitecturas reusables**. Este método propone el aprovechamiento de desarrollos previos, en especial los existentes en la industria. En este grupo Jelliffe, R. (1998) también revisa el concepto de unidades de información propuesto por Maler, E. y El Andaloussi, J. (1996, p156-160) como un tipo de elemento que puede ser entendido y existir por si sólo. Si se identifica que los documentos poseen unidades de información el esquema puede ser diseñado con una arquitectura de microdocumentos, lo cual facilita la administración posterior de la colección (Jelliffe, R. 1998, p. 1-51). Este autor también se refiere a los conceptos de cohesión y acoplamiento de la terminología de ingeniería de software de Constantine y Yourdan (como se cita en Jelliffe, 1998) como herramientas para facilitar el análisis modular de los documentos y la integración de los documentos.
2. **Cascadas y espirales**. Este enfoque es el más común en ingeniería de software y tiene muchos adeptos en el desarrollo de esquemas de SGML, entre ellos:
    a. Maler y El Andaloussi, metodología sumamente estructurada, apropiada para proyectos grandes y nuevos.
    b. Alschuler (como se cita en Jelliffe, 1998, p.1-54), propone un procedimiento más sencillo el cual enfatiza que el enfoque en cascada no siempre es el mejor y la guía del desarrollo la debe dar el uso de los documentos.
    c. Colby, Jackson, et.al. (como se cita en Jelliffe, 1998, p.1-55). Establecen cuatro pasos: definir objetivos, analizar las publicaciones,

modelar el documentos haciendo marcado en muestras reales y validar el esquema contra la sintaxis del lenguaje. Estos mismos autores hacen referencia al uso de metodologías orientadas a objetos para desarrollar lenguajes de marcado, y se basan en que el desarrollo de esquemas contiene elementos de cascada (de lo general a lo específico) y de espiral (iterativo).

d. Travis y Waldt (como se cita en Jelliffe, 1998, p. 1-55). Proponen un modelo menos estructurado y refuerzan la necesidad de iteración.

3. Esquemas exploratorios y prototipos
4. Punto de vista o análisis de escenario

Tanto Jeliffe, R. (1998) como Maler, E. y El Andaloussi, J. (1996) proponen el uso de diagramas como herramienta básica para diseño de esquemas.

Aunque se seleccionó la metodología de las ya mencionadas autoras Maler, E. y El Andaloussi, J. (Idem) se considera que conocer

3.4.2 Metodología de diseño orientado a objetos

Barker, J. (2002) refiere como conceptos básicos de la programación orientada a objetos los siguientes:

1. La abstracción y el modelado, como herramientas de simplificación de la realidad con el fin de hacerla manejable.
2. Los objetos y las clases. Objetos en el ámbito del software son construcciones de software que integran datos y funciones (llamadas métodos) que de forma conjunta representan una abstracción del mundo real. Una clase es un conjunto de objetos que comparte similares características.

3. La interacción entre objetos. Un objeto puede ser puesto en movimiento por una acción externa, puede hacer públicos sus métodos, comunicar sus solicitudes y mantener sus datos resguardados para asegurar la integridad.
4. Las relaciones entre objetos. Pueden existir relaciones estructurales entre clases y entre objetos individuales. Un ejemplo de relación estructural es la herencia, un poderoso mecanismo para derivar nuevas clases que permite indicar que diferencias existen entre las nuevas clases y las ya existentes.
5. Las colecciones de objetos. Son un tipo especial de objetos que sirven para agrupar y organizar otros objetos, y que nos permiten modelar situaciones o conceptos complejos de la vida real.

En el ámbito de las bibliotecas digitales se utiliza el término objeto digital para referirse a un archivo digital parte de la colección digital, en este sentido los objetos representan datos y métodos asociados a un determinado tipo de archivo o clase. Aunque ciertamente no se espera que un archivo tengan todas las características de un objeto dentro de un sistema desarrollado en programación orientada a objetos.

Barker (2002, p. 194) indica que la metodología de desarrollo de sistemas puede ser resumida en 4 pasos:

1. Obtener una descripción narrativa del problema y a partir de ella identificar las diferentes categorías de protagonistas y funciones involucradas.
2. Considerar el aspecto de datos para identificar las clases del mundo real que generarán los objetos de la aplicación y las interrelaciones de los mismos.

3. Considerar la parte funcional identificando como los objetos colaboran para lograr los objetivos del sistema. Aquí se determina que comportamiento será necesario para cada clase.
4. Probar el modelo

3.4.3 Metodología diseño de base de datos

La metodología de diseño de base de datos es mencionada por varios autores (Harold, E.R., 2001 y Jeliffe, R. (1998) como útil en la conceptualización de un diseño de datos a implementar en XML. Elmasri, R. y Navathe, S.B. (2000, cap. 16) proponen dentro del ciclo de vida de un sistema de información los siguientes pasos:

1. Definición del sistema
2. Diseño de Base de Datos
3. Implementación de la Base de Datos
4. Carga o conversión de datos
5. Conversión de aplicaciones previas
6. Pruebas y validación
7. Operación
8. Monitoreo y Mantenimiento

Es de interés en este trabajo el paso 2, el cual los autores ya mencionados traducen en 6 fases

	CONTENIDO Y ESTRUCTURA DE DATOS	APLICACIONES DE BASE
Fase1: RECOLECCIÓN DE REQUERIMIENTOS Y ANÁLISIS	REQUERIMIENTOS DE DATOS	REQUERIMIENTOS DE PROCESAMIENTO
Fase2: DISEÑO CONCEPTUAL DE BASE DE DATOS	DISEÑO CONCEPTUAL DEL ESQUEMA	DISEÑO DE TRANSACC. Y APLICACIONES
Fase3: SELECCIÓN DBMS		
Fase4: MAPEO DEL MODELO DE DATO	MAPA LÓGICO Y DISEÑO DE VISTAS	frecuencias ejecuciones Limitaciones
Fase5: DISEÑO FISICO	DISEÑO INTERNO DEL ESQUEMA	
Fase6: IMPLEMENTACION Y PUESTA A PUNTO	Codificación definición datos Codificación definición almacenamiento	

Figura 3.2 Fases del diseño de bases de datos para grandes bases de datos (tomado de Elmasri y Navathe, 2000, fig.16.1 p. 533)

La segunda y tercera columna de esta figura muestran el enfoque de datos y de aplicación respectivamente. Tradicionalmente los diseñadores de bases de datos se han enfocado a datos y los programadores a procesos, este enfoque está cambiando cada día más y con el auge de la programación orientada a objetos tenemos que ambos elementos se están utilizando conjuntamente para la ingeniería de software.

El XML ha sufrido el proceso equivalente al estar orientado inicialmente a estructuras documentales vistas como partes en un DTD por ejemplo, a enfocarse en los esquemas a estructuras documentales que contienen datos con un potencial de proceso.

Los enfoques del diseño conceptual de bases de datos (fase 2 de acuerdo con

Elmasri, R. y Navathe, S.B. (2000), vendría a ser el elemento más importante para esta investigación y es importante señalar que los autores reportan dos enfoques para el diseño conceptual: el centralizado y el de vista integrada, el primero parte de ver el todo inicialmente y el segundo a partir de vistas parciales ir construyendo uno global.

Con respecto a las estrategias de diseño que revisan estos autores se listan las siguientes:

1. De arriba abajo. Ver el todo y refinar las partes.
2. De abajo a arriba. Partir de lo básico e ir añadiendo o combinando partes.
3. De dentro a afuera. Estilo de arriba abajo que parte de lo evidente y crece con la vecindad de entidades.
4. Combinada

3.4.4 Selección de la metodología de generación de esquema o vocabulario

Para desarrollar la aplicación XML que aquí nos preocupa se decidió utilizar de entre las descritas en el apartado III.4.1 la metodología de Maler, E. y El Andaloussi, J. (1996) por ser la más completa, pensada y desarrollada para el marcado de documentos estructurados y por contar con modelos para la documentación de productos intermedios y finales. En el capítulo 4, se describirá como se generó el vocabulario específico para el marcado de artículos de revista.

Como puede verse en la descripción de los conceptos básicos y de la metodología de desarrollo de sistemas orientada a objetos un diseño de esquema tipo DTD o XSD puede ser desarrollado con esta metodología, pero se considera que si se tiene un proceso propio para este tipo de aplicaciones como la de Maler, E. y El

Andaloussi, J. (1996) es mejor utilizarla y evitar perder eficiencia y perspectiva tratando de adaptar una metodología diseñada con otros objetivos.

No se utilizó tampoco la metodología de diseño de bases de datos durante la presente investigación por razones similares a las que se desecho la metodología de diseño orientado a objetos; es decir, por no ser específica para este tipo de aplicaciones. Sin embargo, se retomará la importancia de considerar tanto datos y procesos en un mano a mano tal y como lo señala la metodología de diseño de bases de datos al considerar los elementos presentes en la DTD siempre relacionados con la búsqueda de información y el despliegue de la misma.

3.5 Buscador

La selección del software buscador para llevar a cabo las pruebas fue considerada en virtud de las siguientes características:

- Software libre
- Que permitiera aprovechar el marcado (este punto se volvió eventualmente menos fundamental en la medida que la colección de artículos era pequeña y la relación marcado y relevancia podía hacerse de forma manual)
- Conocido por instituciones o individuos con los que se pudiera tener contacto para evitar retrasos innecesarios en caso de dificultades técnicas.

Una evaluación de buscadores llevada a cabo por Morgan, E.L.(2001) sobre buscadores de software libre indicaba que Harvest cumplía con el primer requerimiento.

Una consideración importante al seleccionar Harvest[8] como el buscador para llevar a cabo la búsqueda de información de los documentos con y sin marcado es que este buscador es un software libre que permite el aprovechamiento o no del marcado de documentos.

La segunda consideración fue que la UNAM-DGSCA lo está utilizando como buscador que aprovecha el marcado XML de textos y esa referencia facilitaba la consulta técnica en caso de problemas relacionados con la instalación del software.

Aspectos relevantes de la implementación se verán en el capítulo 4.

3.6 Evaluación de resultados de búsqueda

La evaluación en general de los sistemas ha sido un tema muy estudiado, Buttenfield, B. (1999, p.40) reporta una bibliografía sobre estos enfoques que cubre necesidades y evaluación de los usuarios, avances etnográficos y empíricos, y progreso tecnológico en paralelo. La pregunta de si un sistema de recuperación de información es bueno ha estado vigente entre la comunidad relacionada con el cómputo y las bibliotecas, y Borgman, C. L. (2000) lo reporta como uno de los temas más recurrentes en las en las conferencias de la CHI (Computer Human Interface) de la ACM (Association of Computer Machine).

Lancaster, F.W. (1979, p.198) considera que los factores a evaluar para determinar si un sistema de recuperación de información son costo, tiempo y calidad de los resultados. Para esta investigación el enfoque será en el aspecto calidad de resultados a través de la precisión de los mismos.

[8] http://harvest.sourceforge.net

El tema que aquí nos interesa es el de la evaluación de los resultados de una búsqueda ha sido explorado desde 1955 (Borgman, C.L., Moghdam, D. y Corbett, P.K., 1984, p. 134) cuando Kent, et.al. proponen una serie de medidas sobre la calidad de los resultados de las cuales, ya para 1984 sólo se utilizaban dos[9]:

1. "Precisión": llamado también el factor de pertinencia (Idem, p.134) y relevancia del resultado (Lancaster, F.W. , 1979, p.109). Es la tasa de registros relevantes recuperados en relación a todos los registros recuperados, se expresa como una fracción:

$$\text{Precisión: } = \frac{\text{Registros relevantes recuperados}}{\text{Todos los registros recuperados}}$$

2. Recall: completeness of the output (Idem, p.109), relación de recuperación (Lancaster, F.W., 1983, p.162). Es la tasa registros recuperados en relacion a todos los registros relevantes en la base de datos, se representa con la siguiente fracción:

$$\text{Relación de recuperación} = \frac{\text{Registros relevantes recuperados}}{\text{Registros relevantes en la base de datos}}$$

Borgman, C.L., Moghdam, D. y Corbett, P.K., (1984, p.135) señalan como estas medidas están basadas en búsquedas individuales más que en la ejecución general del sistema; establecen además las limitaciones que conllevan estas tasas, a saber:

[9] Los términos pertinencia y relevancia son utilizados por algunos autores indistintamente para referirse a un ítem que es útil a un usuario para satisfacer sus necesidades de información. En este texto se utilizará relevancia para referirse a ese concepto.

1. Dependen de la cobertura de la base de datos (en este caso de la colección).
2. El concepto de registros relevantes es totalmente subjetivo, un mismo individuo puede variar su juicio de un momento a otro.
3. Para calcular el Recall o relación de recuperación no es sencillo, sobre todo en la medida que una colección es grande. En el caso de grandes repositorios se han propuesto métodos como muestreo y búsqueda exhaustiva de un tema.

A pesar de las limitaciones expuestas, autores preocupados por el tema de la evaluación de los resultados de búsqueda desde distintas disciplinas siguen utilizando estas medidas Lancaster, F.W., (1979, 1983 y 1993); Borgman, C.L. (1984 y 2000), Stern, D. (1999) en bibliotecología y Witten, I.H, (1999, 2003), Belew, R. K (2000) y Fuhr, N. et. al. (2001) en computación. Y aunque específicamente Fuhr, N. et. al. (2001, p. 188) considera que una visión más amplia podría enfocarse en tiempo dedicado a la tarea o tasa de “completion” sigue tomando como válidas las medidas descritas.

Se debe aclarar que las medidas de precisión y relación de recuperación (recall) fueron diseñadas para bases de datos bibliográficas y no para repositorios de texto completo, sin embargo el hecho de utilizarlas en el ámbito del texto completo sigue siendo válido pues nos indica por un juez experto si el material es útil o no con respecto al tema con el beneficio de que el juicio ya no se basa en una ficha bibliográfica con o sin resumen, sino en el texto completo de artículo.

Como se mencionó en el apartado III.2 existe una iniciativa (INEX) que pretende medir el impacto del marcado XML en la recuperación de información (Baeza-

Yates, R., Fuhr, N.Y. Maarek, 2002 y Govert, N. y Kazai, G., 2003) , la metodología utilizada por este proyecto nacido del grupo de interés sobre recuperación de información (IR) de la ACM incluye:

- Participantes con buscadores enfocados a IR, a bases de datos y propios de XML.
- Búsquedas sobre el texto completo llamadas de contenido y sobre textos marcados llamadas de contenido y estructura.
- Participaron 49 organizaciones de 21 países con una base de datos de 12, 107 artículos correspondientes a 12 revistas y 6 transacciones de la IEEE en el tema de cómputo.
- Se midió precisión y relación de recuperación (recall) utilizando multiescalas de 4 niveles en cada caso.
- Los resultados generales fueron en promedio que para las búsquedas con contenido y estructura la precisión llegó a un máximo de 34% y en las búsquedas en texto completo hasta el 27%
- Aunque las búsquedas se llevaron a cabo utilizando el texto completo se asignaron temas y el desarrollo de los esquemas temáticas fue parte importante del ejercicio.

Dado que el interés principal de INEX es probar la eficiencia de los buscadores y la forma en que aprovechan el XML se tomó el marcado DTD utilizado por IEEE sin discusión de los elementos que ciertamente son bastante adecuados, aunque puede considerarse que no explotan todas las posibilidades de contenido (Idem, pp. 4-8)

Entre esta investigación y el INEX existen similitudes, pero la diferencia más importante es que en el caso de esta tesis el centro de la investigación es el desarrollo del vocabulario e INEX asume no tener incidencia en el marcado y se enfoca en aprovechar lo disponible. Se espera conocer los resultados del INEX 2003 y que se pueda colaborar con ese proyecto el próximo año con el marco de referencia de este estudio.

Este capítulo registró aspectos metodológicos en relación a la investigación en sus facetas:

- Descriptivas: diseño e implementación del vocabulario
- Exploratoria: medición de resultados de precisión y su correspondiente preexperimento.

En el siguiente capítulo se describirá el diseño e implementación del vocabulario propuesto, la creación de la colección de documentos para la prueba, la instalación del buscador harvest y la recolección de los datos de la evaluación

4. VOCABULARIO XML PARA ARTÍCULOS DE REVISTA Y SU IMPACTO EN LA PRECISIÓN DE LOS RESULTADOS DE BÚSQUEDA.

En el capítulo tercero se revisó la metodología para llevar a cabo una investigación exploratoria que indique una tendencia en relación a la hipótesis propuesta:

> *Al marcar artículos de revista con el vocabulario XML adecuado los resultados de la búsqueda sobre ciertas partes marcadas del texto de los artículos serán más precisos que la misma búsqueda sobre el texto completo.*

La concreción de la metodología expuesta en el capítulo anterior se describe en este capítulo cuarto, cuyos contenidos son:

- El diseño e implementación de un vocabulario XML para marcar artículos de revista
- La identificación, instalación y configuración de un buscador de software libre que permitió llevar a cabo el preexperimento descrito en el apartado 3.3
- La recolección y análisis de los datos referidos a la evaluación de la precisión en los resultados de búsqueda, de acuerdo al diseño experimental.

4.1 Diseño e implementación del vocabulario

Como se indicó en el Capítulo 3 la selección de la metodología de desarrollo para el vocabulario XML a utilizar en el marcado de artículos de revista fue la de Maler, E. y El Andaloussi, J. (1996), por considerarla la más completa entre las disponibles para el desarrollo de DTD para texto con cierta estructura. Cabe

aclarar que la mencionada metodología podría adaptarse a contenidos con estructura fija más orientados a los datos a implementarse en esquemas más modernos como el XSD.

La metodología seleccionada es bastante estructurada y está basada en el modelo llamado de cascada (de lo general a lo específico). Sin embargo, en el momento real de la implementación, como todo diseño, requiere que en algunos casos se revisen algunos de los pasos; estas iteraciones están indicadas en la metodología misma que no escatima en detalles para asegurar el éxito del proyecto de desarrollo del vocabulario[1].

Previo a la metodología del diseño, las autoras mencionadas, consideran necesario que el equipo de desarrollo se familiarice con los diagramas de árbol y con los conceptos básicos del lenguaje de marcado, sin establecer demasiados detalles de la sintaxis del DTD para evitar que el diseñador se aboque a generar código antes de tiempo. Maler y El Andaloussi dedican en su estilo minucioso, antes de la etapa de diseño una etapa de administración del proyecto, que no será considerada en este trabajo.

4.1.1 Diseño del vocabulario

Para el diseño del vocabulario se llevaron a cabo 10 pasos integrados en las 6 fases expuestas en el apartado 3.4.1

Fase 1: Articular los objetivos del proyecto

El objetivo del vocabulario en este caso está en sintonía con el objetivo de la investigación:

[1] Es muy probable que algunos diseñadores de aplicaciones consideren esta metodología excesivamente guíada, en esta investigación el detalle mismo fue de gran ayuda.

Proponer y evaluar un vocabulario XML que describa y estructure adecuadamente los contenidos de artículos de revistas académicas en formato digital, dicha vocabulario facilitará la diversidad de representaciones de tales objetos digitales y la recuperación de información de los mismos.

Durante todo el diseño la idea guía ha sido que el marcado facilite la recuperación de la información esencial de los textos. Cabe aclarar que se ha tomado en cuenta que el marcado debe facilitar también el despliegue para potenciar la entrega de información al usuario.

Fase 2: Análisis de las necesidades que pretenden satisfacer los datos contenidos en los documentos

Paso 1: Identificar y definir los componentes de información básica

Un componente será una pieza del texto que puede reflejar información referente a contenido, estructura o presentación gráfica (Idem, pp.95-96). En este caso los componentes que guiarán el desarrollo del vocabulario serán los de tipo contenido, aunque como se dijo antes, no puede obviarse los otros dos tipos.

Hay que aclarar que los componentes de esta primera lista (Tabla 4.1) se generaron con base a la experiencia en el uso y producción de artículos de revista, y a través de consultas[2] con otros editores e investigadores usuarios de repositorios de artículos de revista en formato electrónicos y cotidianos lectores de

[2] Estas consultas se llevaron a cabo de forma casual y permitiendo al entrevistado opinar abiertamente sobre la búsqueda y recuperación de información, ya que los individuos tienen interés no solo en opinar sobre estructuras de artículos de revista, sino también sobre sus expectativas y experiencias con los sistemas de recuperación de información.

revistas académicas en papel. Entre los comentarios/solicitudes más importantes tenemos:

- Que los interfases deben ser lo más sencillos posibles
- Que se pudiera buscar exclusivamente en los encabezados dentro del artículo
- Que en listas de resultados con muchos ítems, aquellos en donde apareciesen las palabras presentes en ciertas partes de la estructura del artículo aparecieran antes.
- Que no le gustaría tener dos interfases.
- Que no se obligará a aprender nuevas sintaxis o formas avanzadas de búsqueda.

Los comentarios anteriores no necesariamente se enfocan a la indentificación de componentes de información básica, pero son valiosos en la medida que nos indican las expectativas de los usuarios con respecto a los sistemas de recuperación de información

Titulo	Conclusiones	negritas
Autor	Bibliografía	italicas
Institución del autor	Agradecimientos (personas y/o instituciones)	centrado
Palabras Claves	párrafo	tamaño fuente
Resumen	figura	color
Introducción	lista	párrafo sangrado
Div. 1er Nivel	tabla	
Div. 2º Nivel	nota	
Titulo Div. 1er Nivel	referencia bibliográfica	

Tabla 4.1: Lista componentes básicos.

Para cada componente se llenó una Forma de Análisis de Componentes. En la Tabla 4.1 se listan todos los componentes analizados en esta fase inicial :

Forma de Análisis de Componentes

Nombre del elemento: Título _____ Número:_____

Definición: Nombre, frase que contiene una referencia más o menos explicativa de la materia o argumento de un objeto digital textual. Incluye subtitulo, títulos paralelos y cualquier frase que pueda considerarse parte del título. _____
_____

Clases: Contenido y Descriptivo. DC: Title_____

Explicación y ejemplos: Se respeta el titulo que el autor ha dado a la obra aunque no logre describirla adecuadamente.

Marcado existente: Centrado_____ _____
Tipo más grande _____
_____ _____

_ _

Aceptado: _____

Justificación: Todo artículo de revista posee un titulo_____

Elementos de marcado XML relacionados: _____
_____

Historia de creación y cambio: 24/03/2003 se creo y 25/03/2003 se buscó equivalente en DC

Figura 4.1 Forma para el análisis de componentes. Ejemplo componente título.

Paso 2: Clasificar los componentes en grupos lógicos

Una vez generado el listado inicial se clasificaron los componentes en:

Componentes Descriptivos (contenido fuera del texto principal)
Titulo
Autor
Institución del autor
Componentes de Contenido
Palabras Claves
Resumen
Introducción
Div. 1er Nivel
Título Div. 1er Nivel
Div. 2º Nivel
Título Div. 2º Nivel
Conclusiones
Bibliografía
Agradecimientos (personas y/o instituciones)
Componentes Estructurales
Párrafo
Figura
Lista
Tabla
Nota
Referencia Bibliográfica
Componentes de Presentación Gráfica
Negritas
Itálicas
Centrado
Tamaño fuente
Color
Párrafo sangrado

Tabla 4.2 Componentes Clasificados

Un punto a señalar es que es difícil separar formato de contenido como en el caso de negritas e itálicas (A very gente ...SGML is a “declarative mark-up language”, parr.20); o estructura y contenido como en el caso de textos que representan el título de una división del artículo. En ambos casos el mismo componente de presentación gráfica o de estructura puede incluir palabras que reflejan de forma importante el contenido.

Paso 3: Validar el análisis contra otros modelos existentes

Tal y como se estableció en el apartado II.4.6 existen esquemas como MARC, DC y SciELO (Scientific Electronic Library Online) enfocados a los metadatos descriptivos y de los cuales se han desarrollado implementaciones utilizando XML.

Los diseños que más interesaron como referencia para esta investigación son los de marcado estructural como el TEI, y los que están enfocados a artículos de revista como el JAI (Journal Archiving and Interchange)-JP (Journal Publishing) y SciELO. En el caso de TEI, debido a la extensa cantidad de elementos propuestos en el marcado completo (450 elementos), se estudió la versión llamada TEI Lite (150 elementos) y una versión pedagógica llamada texbaby (60 elementos)[3]

SCIELO	TEI Lite	J.Arch.& Interchange (JAI)
Article	TEI.2	Article
front-F	teiHeader-Th	Front-F
body-Bo	text-T	Body-Bo
back-Ba		Back-Ba
		Sub-Article-S
		Response-R
F-titlegrp	Th-fileDesc	F-metadata
F-authgrp	Th-profileDesc	F-authorNoteGroup
F-bibcom	Th-revisionDesc	F-articleGropingData
Ba-bbibcom	T-front	F-TitleGroup
Ba-vancouv	T-body	F-ContributorGroup
Ba-iso690	T-back	F-Contributor
Ba-abnt6023		F-PublicationDate
Ba-other		F-DocumentHistory
		F-Abstract
		F-Conference
		F-Notes
		Bo-Section
		Ba-AppendixMatter
		Ba-Appendix
		Ba-GlossElementsList
		Ba-GlossaryGroup
		Ba-BibliograpRefList
		Ba-NMLCitationModel
		Ba-PerGroupCitPublic.

Tabla 4.3 Tabla con los elementos raíz y dos primeros niveles de de los esquemas SCielo, TEI Lite y Journal Archiving Innitiative

[3] http://bistro.northwestern.edu/mmueller/ariadne/babytei/teixbaby.txt

La Tabla 4.3 muestra los elementos raíz y los 2 primeros niveles de estructura de los tres esquemas que más influyeron en esta investigación, y que impactaron en importantes decisiones de diseño. Los puntos más sobresalientes del estudio de los vocabularios fueron:

- SciELO resultó ser una lenguaje de marcado con gran desarrollo a nivel de elementos bibliográficos y de portada, tan específico que podía volverse engorroso[4], pero importante para el análisis bibliométrico[5] ; además del DTD la metodología SciELO facilita todo un esquema para edición, administración de archivos y publicación de los mismos basado en software libre.
- JAI resultó ser fundamentalmente simple, pero con la pérdida del potencial en recuperación de contenido inmerso en el texto.
- Dado que el TEI es el vocabulario que más desarrolla componentes a nivel estructural el principal vocabulario de comparación sería TEI y en la medida de lo posible se trataría de mantener o traducir las etiquetas a fin de apegarse a un estándar (Bournard, L., 2000, junio). El TEI, ciertamente presentó dificultades, no fue fácil encontrar ejemplos de marcado en revistas y aun las versiones mas sencillas poseen complejidad, afortunadamente menos en la parte correspondiente al cuerpo del texto que era la que más interesaba. Después de una ardua búsqueda se identificó

[4] Se asistió a un curso de Metodología SciELO impartido en la UNAM durante mayo de 2003 con capacitadotas provenientes de OPS Brasil y CONACYT Chile respectivamente, lo cual resultó de gran valor durante esta investigación.
[5] Bibliometría: estudio estadístico de las publicaciones en relación a sus diversos aspectos; uno de los más reportados, como en el caso de SciELO, son las citas bibliográficas.

que en la Biblioteca Digital de Cervantes Virtual se utilizaba el TEI para el marcado de todos sus materiales, incluyendo publicaciones periódicas.

- El vocabulario propuesto no pretendería desarrollar metadatos descriptivos en gran detalle pues ya existe mucho aporte previo en relación a ello, desde el formato MARC hasta el Dublín Core, para descripción de tipo catalográfico hasta los componentes presentes en todos los vocabularios como el TeiHeader del TEI, y los "Front's" tanto de SciELO como de JAI.
- No se tomó TEI como modelo exacto ya que aunque enfocado a la estructura no considera todos componentes de contenido que esta tesis pretende.
- Las etiquetas fueron en español de forma que la característica de XML de ser legible a los humanos se mantuviera lo más vigente posible para población de habla hispana.[6]

En este paso parte del análisis consistió en marcar 2 artículos con el TEI a fin de experimentar la dificultar y las bondades en relación a la recuperación de contenidos. Se marcaron también 2 artículos con etiquetas de SciELO como parte del taller al que se asistió. Este ejercicio fue de gran utilidad pues permitió experimentar con el marcado en artículos reales y en un desarrollo integral como

[6] La necesidad de establecer un vocabulario con las características de describir adecuadamente materiales digitales y con etiquetas en español fue el origen de esta investigación, la cual por limitaciones ha tenido que enfocarse a artículos de revista. Esas ideas iniciales fueron propuestas por la Mtra. Clara López, directora de este trabajo, desde el año 2000. Lamentablemente las pruebas de investigación relacionada con la utilización de etiquetas en español no tuvieron cabida en esta tesis por falta de tiempo. Hoy día están publicadas algunas propuestas de multilingualidad en el marcado TEI como las de Bia, A.G., Sánchez-Quero, M. y Deau, R. (2003, mayo). Como nota anecdótica personalmente comenté con Alejandro Bía este interés durante la reunión de Interfases, Colima noviembre 2001, afortunadamente él tuvo el tiempo y recursos para llevar a cabo propuestas al respecto que deben difundirse y continuarse para ir desarrollando un lenguaje de marcado estructural común.

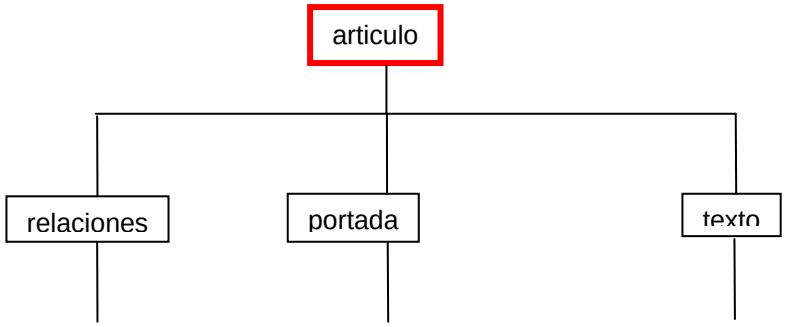
ya comentó en el párrafo anterior. Esta fase del proceso de diseño también ayudó a decidir que el vocabulario propuesto en esta tesis sería original y un potencial espacio de nombre y no un perfil de aplicación.

Fase 3: Modelado: Diseñar los requerimientos del esquema basándose en los objetivos.

Una vez llevadas a cabo las dos primeras fases: establecimiento de objetivos y análisis de las necesidades de información que los textos deben y pueden satisfacer, se pasó de lleno al modelado. Durante los diferentes pasos que conlleva esta fase se utilizaron los diagramas de árbol y una forma parecida a la de análisis de componentes (Fig. 4.1) llamada de Forma de Elementos.

Forma de Elementos

Nombre del elemento: articulo_____ Número:01_____
Identificador Genérico:_____
Clase(s): _____
Modelo:



Contenido: Es el elemento raiz o root element y contiene todo la información del documento
_____
_____
Justificacion: Debe existir un elemento raiz y articulo es un buen nombre pues cada archivo contiene un articulo
_____
_____

Es contenido en: _____

Componentes relacionados: relaciones, portada y texto _____
_____

Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 diagrama _____
_____
_____
_____

Figura 4.2 Forma de Elementos.

Paso 4: Seleccionar los componentes semánticos

Los componentes originalmente expuestos en el paso en las Tablas 4.1 y 4.2 fueron enriquecidos con los componentes presentes en los vocabularios TEI, SciELO y JAI-JP, y a través del marcado de artículos reales se generó una nueva lista de componentes

articulo	div0 hasta div5	encabezado	**p**
relaciones	**figura**	enfatizado	
portada	**lista**	otroIdioma	
texto	**tabla**	lugar	
	bibliografía	siglas	
ejemplar	agradecimiento	fecha	
revista	descFig	ejemplo	
fuente	item	formula	
autor	**itemBib**	definicion	
titulo	fila	termino	
fechaRecibido	celda	**palabreClave**	
fechaAceptado	nombre	cita	
epígrafe	cargo	**nota**	
resumen	**institución**	liga	
	mail	citaText	

Tabla 4.4 Lista de Elementos del vocabulario articulo. En negritas los componentes rescatados de la lista original en Tabla 4.2.

Este listado pasó de ser bastante general a nivel de componentes (Ver tablas 4.1 y 4.2) a contener más elementos estructurales y de contenido que se irían afinando en los pasos restantes.

Paso 5: Construir la jerarquía global del documento y la meta información.

Una vez se tuvo el listado con la mayoría de los elementos propuestos se pasó a la esquematización de la jerarquía. En realidad para ir evaluando la estructura y los elementos mismos fue necesario elaborar diagramas para visualizar el contenido y sus relaciones generales. Los diagramas se elaboraron con la metodología de árboles de acuerdo a las indicaciones de Maler, E. y El Andaloussi, J. (1996), como puede verse en la figura 4.3.

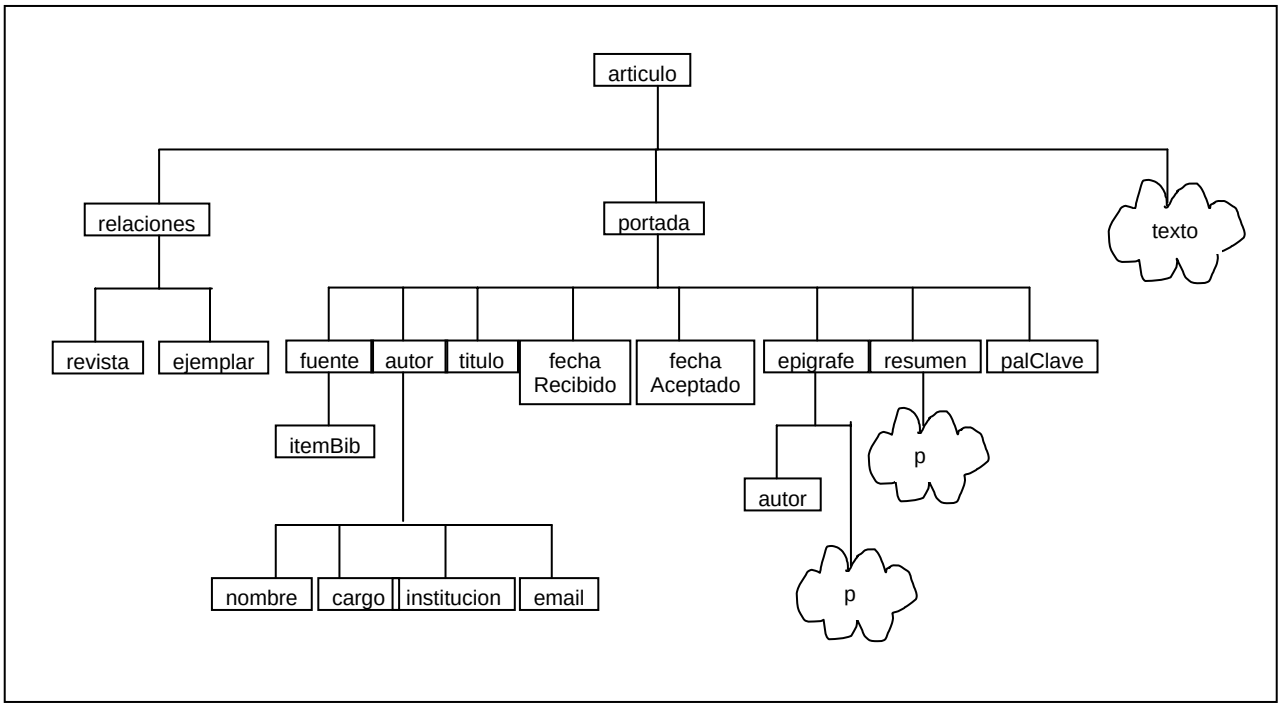


Figura.4.3. Vocabulario articulo, jerarquía global y desarrollo de los elementos relaciones y portada

La jerarquía global se compone de un elemento raíz necesario en la especificación de un vocabulario de XML y 3 elementos principales:

- relaciones: establece las ligas con el título de revista y con el ejemplar, que se pueden describir en el documento o ser entidades relacionadas
- portada: en donde se vacían los metadatos descriptivos referentes al documento
- texto que engloba el texto principal y por lo tanto el elemento más importante para este vocabulario.

Paso 6: Construir el modelo de elementos y atributos a nivel medio del documento: unidades de información

Una vez establecidos los elementos y estructura general se avanzó a las estructuras del nivel medio del documento. Los diagramas de árbol correspondientes pueden verse en el Anexo 1. Diagrama de árbol del Vocabulario artículo; por su estructura vertical y por las formas que conlleva dichos diagramas

son un poco engorrosos al ir creciendo las ramas. En este apartado se utilizarán los diagramas de árbol horizontales, utilizando texto y flechas, los cuales son más sencillos comunes en la documentación de DTD's como puede verse en el sitio Web de SciELO o JAI-JP.

Al visualizar las unidades de información que irán conformando la estructura e ir estableciendo si los elementos llevarán o no atributos (descritos en el apartado 2.4.1) y las especificaciones de ocurrencia. Las ocurrencias se indican de la siguiente manera:

- Si el elemento es obligatorio no se indica ningún símbolo
- Si el elemento es opcional se utiliza el símbolo ?
- Si el elemento es opcional y puede ser repetible se utiliza el símbolo *
- Si el elemento debe aparecer al menos una vez y puede ser repetible se utiliza el símbolo +

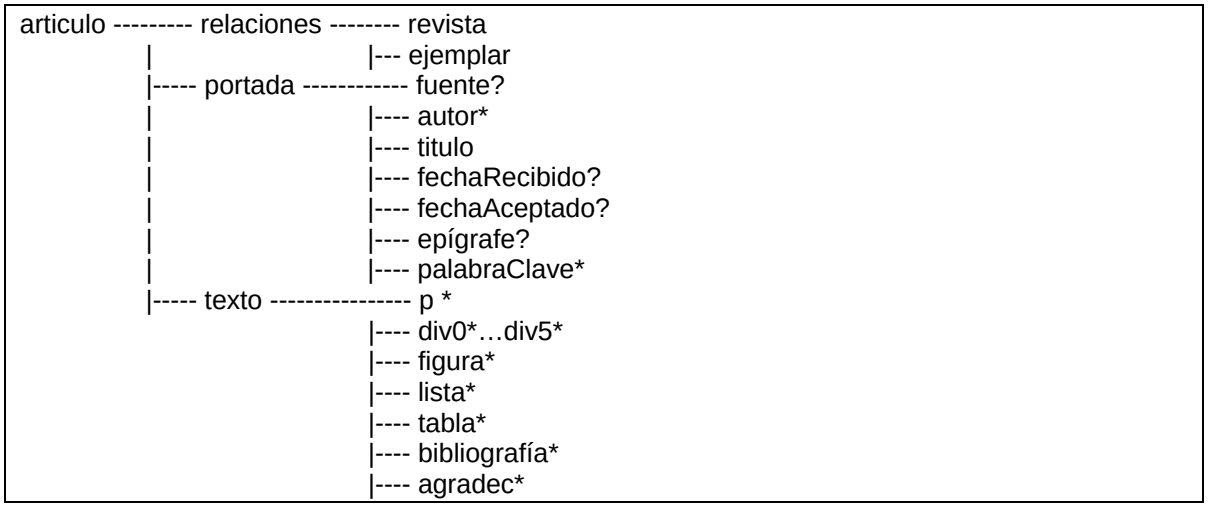


Figura 4.4 Elementos y especificaciones de ocurrencia a nivel medio.

Elementos tales como **revista, ejemplar, autor, div0** hasta **div5, figura, lista** pueden considerarse unidades de información, de tal forma que si se quisiera desarrollar un esquema con entidades coordinadas se podrían indicar ID en

algunos de ellos como revista, ejemplar y autor para desarrollar un sistema de información.

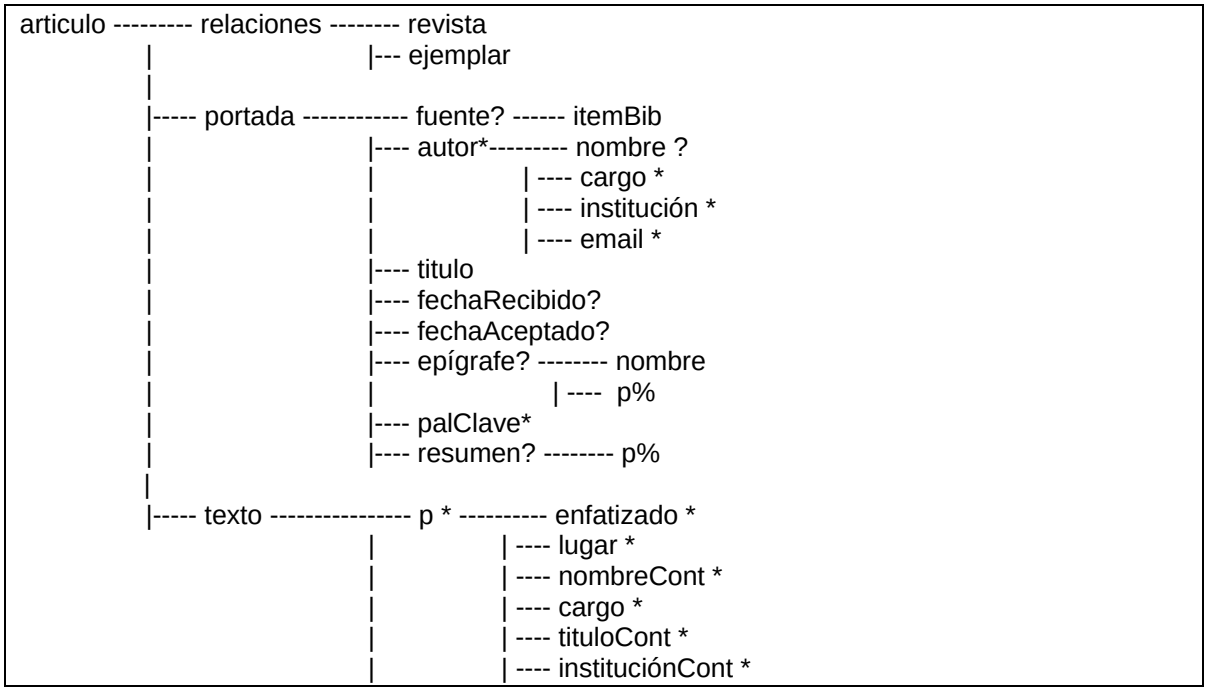
Se decidió en el diseño que los elementos podrían llevar como atributos generales la tipografía y un número en caso de referencia a otras entidades relacionadas.

Los elementos **div0, div1, div2, div3, div4, div5** se consideran como divisiones del texto que pueden de forma anidada, es decir, si la **div0** es un apartado general, **div1** serían los apartado integrados en **div0**, y así sucesivamente.

Algunas unidades de información tienen ubicación en diferentes partes de la estructura, por ejemplo, el elemento **figura** puede ubicarse directamente bajo texto o bajo los elementos **div0** y **p**.

Paso7: Construir el modelo de elementos y atributos para jerarquía de bajo nivel del documento o elementos a nivel de datos

Este paso establece la ubicación estructural de los elementos que llevarán en su mayoría los contenidos textuales.



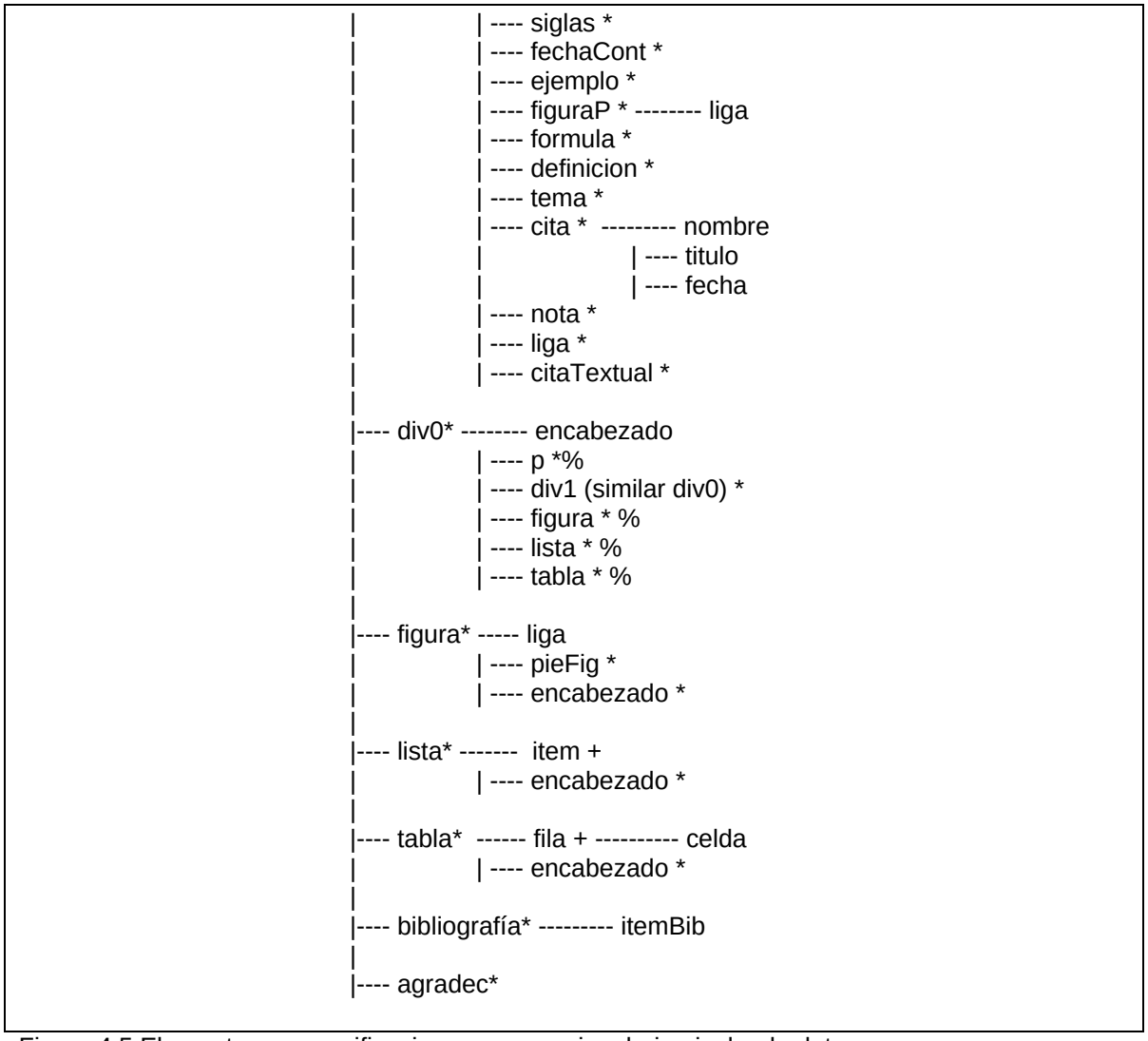


Figura 4.5 Elementos y especificaciones ocurrencia a bajo nivel o de datos

Los signos de % representan unidades de información que tienen elementos hijos descritos en otras partes del diagrama.

Paso 8: Proponer variedad de elementos de donde un autor pueda seleccionar los necesarios, poblar las ramas que hayan quedado generales

En la tabla 4.5 se muestra una matriz a nivel de datos en donde al establecer los contextos de los diversos contenidos facilita el desarrollo de las ramas del árbol de la estructura de marcado.

Contenido/contexto	portada	portada/autor	texto	texto/p	texto/p/cita	texto/div0
Nombre		x		X	x	
Titulo	x			X	x	
institución		x		X		
Figura			x	X		X

Tabla 4.5 Distribución de algunos contenidos en la estructura de marcado.

Paso 9:Hacer conexiones entre el modelo y el mundo externo

Se considero a nivel de diseño que las conexiones internas y externas de este modelo son:

- Las entidades internas del sistema de forma tal que **artículo**, **revista** y **ejemplar** sean DTD´s independientes, esto a nivel de diseño pues en la implementación se simplificará ya que no impacta en la comprobación de la hipótesis de investigación.
- Las imágenes se llamarían a través de ligas, lo referente a ligas tampoco tiene un impacto en la investigación.
- La publicación electrónica de los documentos será a través de Internet por lo que será necesario crear XSL para la conversión de los archivos marcados con XML en versiones HTML accesibles vía web. Este punto de diseño si tiene mucho impacto para la evaluación del vocabulario por lo que se implementará.

Paso 10: Validar que el modelo esté completo y que está tomando en cuenta otros modelos similares ya desarrollados

Una segunda revisión del vocabulario TEI, SciELO y JAI-JP y del marcado de artículos de revista de la IEEE utilizados en las pruebas INEX (Govert, N. y Kazai, G. 2003, pp. 4-5) confirmó la decisión inicial de que el esquema TEI era lo más

cercano a los objetivos de la investigación, SciELO no marcaba estructuralmente, JAI-JP e IEEE consideraban la estructura pero sólo a nivel grueso de división. TEI no se tomó como vocabulario origen sino como base para el propio diseño.

4.1.2 Implementación del vocabulario

Fase 4: **Completar el diseño del esquema e implementarlo.**

El diseño se implementó a través de un DTD, se exploró la implementación a través de un XSD pero dada las características de los textos se puso de manifiesto que la versatilidad necesaria para esta marcado posible en un DTD no lo era con XSD. Por ejemplo al declarar el elemento autor

```
<!ELEMENT autor (#PCDATA | nombre | cargo | institucion | email)*>
```

es posible que los elementos hijos se presenten en cualquier orden dentro de la etiqueta autor y en las repeticiones que van desde 0 hasta muchas, en XSD el orden no puede obviarse y el número de repeticiones debe indicarse, esto último se debe a que XSD está orientado a datos y no a texto. A pesar de tratarse de una sintaxis con bastantes años la vigencia de los DTD es actual en la medida que es tanto la sintaxis de esquema del SGML como la original de XML y se adapta a las características de los textos narrativos.

Cabe destacar que:

- Se consideraron inicialmente atributos generales en el DTD; sin embargo al llevar a cabo el marcado se consideró que no era necesario implementarlo ya que no tenían impacto en las pruebas de recuperación de información.
- Como se indicó en el paso 9. Se incluyeron a nivel de diseño las entidades externas revista y ejemplar, pero se dejaron a ese nivel ya que tampoco

tenía mucho sentido desarrollar elementos de metadatos descriptivos cuando lo más importante era el marcado del elemento texto.

Ambas consideraciones, aunque no implementadas en esta investigación, pueden ser implementadas posteriormente.

La figura 4.6 muestra el código completo de la DTD que se utilizó para validar el marcado de los documentos.

```
<!ENTITY % atributos-generales "tipografia CDATA #IMPLIED
                                numero CDATA #IMPLIED">
<!ELEMENT articulo.1 (relaciones, portada, texto)>
<!ELEMENT relaciones (revista, ejemplar)>
<!— revista y ejemplar como entidades externas
<!ENTITY % revista SYSTEM "revista.dtd">
<!ENTITY % ejemplar SYSTEM "ejemplar.dtd"> -->
<!— entidades necesarias para el reconocimientos de caracteres en español
<!ENTITY % symbol SYSTEM "HTMLsymbol.ent">
%symbol;
<!ENTITY % lat1 SYSTEM "HTMLlat1.ent">
%lat1; -->

<!ELEMENT revista (#PCDATA)>
<!ELEMENT ejemplar (#PCDATA)>
<!ELEMENT portada (autor+, titulo, fuenteBib?, fechaRecibido?, fechaAceptado?, epigrafe?,
resumen?, palClave*)>
<!ELEMENT fuente (itemBib)>
<!ELEMENT autor (#PCDATA | nombre | cargo | institucion | email)*>
<!ELEMENT nombre (#PCDATA)>
<!ELEMENT cargo (#PCDATA)>
<!ELEMENT institucion (#PCDATA)>
<!ELEMENT email (#PCDATA)>
<!ELEMENT titulo (#PCDATA)>
<!ATTLIST titulo
        %atributos-generales;
>
<!ELEMENT fechaRecibido (#PCDATA)>
<!ELEMENT fechaAceptado (#PCDATA)>
<!ELEMENT epigrafe (#PCDATA | p | autor)*>
<!ELEMENT resumen (p)*>
<!ELEMENT palClave (#PCDATA)>
<!ELEMENT texto (#PCDATA | p | div0 | div1 | div2 | div3 | div4 | resumen | figura | lista | tabla |
bibliografia | notas | agradec)*>
<!ELEMENT p (#PCDATA | enfatizado | otroIdioma | lugar | nombreCont | cargo | tituloCont |
institucionCont | siglas | fechaCont | ejemplo | formula | definicion | tema | cita | autorCit | nota | liga
| citaText)*>
<!ELEMENT enfatizado (#PCDATA | p)*>
<!ATTLIST enfatizado
        %atributos-generales;
>
<!ELEMENT otroIdioma (#PCDATA | p)*>
```

```
<!ELEMENT lugar (#PCDATA)>
<!ELEMENT nombreCont (#PCDATA)>
<!ELEMENT tituloCont (#PCDATA)>
<!ELEMENT insititucionCont (#PCDATA)>
<!ELEMENT siglas (#PCDATA)>
<!ELEMENT fechaCont (#PCDATA)>
<!ELEMENT ejemplo (#PCDATA)>
<!ELEMENT formula (#PCDATA)>
<!ELEMENT definicion (#PCDATA | p)*>
<!ELEMENT tema (#PCDATA)>
<!ELEMENT cita (#PCDATA | autorCit | titCit | fechaCit | pagCit)*>
<!ELEMENT autorCit (#PCDATA)>
<!ELEMENT titCit (#PCDATA)>
<!ELEMENT fechaCit (#PCDATA)>
<!ELEMENT pagCit (#PCDATA)>
<!ELEMENT nota (#PCDATA | p)*>
<!ELEMENT liga (#PCDATA)>
<!ELEMENT citaText (#PCDATA | autorCit | fechaCit | pagCit)*>
<!ELEMENT div0 (#PCDATA | div1 | encabezado | p | lista | figura | tabla | resumen)*>
<!ATTLIST div0
        %atributos-generales;
>
<!ELEMENT div1 (#PCDATA | div2 | encabezado | p | lista | figura | tabla | resumen)*>
<!ELEMENT div2 (#PCDATA | div3 | encabezado | p | lista | figura | tabla | resumen)*>
<!ELEMENT div3 (#PCDATA | div4 | encabezado | p | lista | figura | tabla | resumen)*>
<!ELEMENT div4 (#PCDATA | div5 | encabezado | p | lista | figura | tabla | resumen)*>
<!ELEMENT div5 (#PCDATA | encabezado | p | lista | figura | tabla | resumen)*>
<!ELEMENT encabezado (#PCDATA)>
<!ATTLIST encabezado
        %atributos-generales;
>
<!ELEMENT figura (#PCDATA | liga | encabezado | pieFig)*>
<!ELEMENT pieFig (p)>
<!ELEMENT lista (#PCDATA | item | encabezado)*>
<!ELEMENT item (#PCDATA | p | lista)*>
<!ELEMENT tabla (fila*, encabezado?)>
<!ELEMENT fila (celda)*>
<!ELEMENT celda (#PCDATA | p)*>
<!ELEMENT bibliografia (#PCDATA | itemBib)*>
<!ELEMENT itemBib (#PCDATA)>
<!ELEMENT notas (#PCDATA)>
<!ELEMENT agradec (p)>
```

Figura 4.6 DTD para el vocabulario articulo

Un ejemplo de los artículos que se marcaron se puede ver en la figura 4.7

```
?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="articulo.xsl"?>
<!DOCTYPE articulo.1 SYSTEM "articulo1.dtd">
<articulo.1>
        <relaciones>
                <revista>Pedagogia</revista>
                <ejemplar>vol 11</ejemplar>
        </relaciones>
        <portada>
```

```
                <autor><nombre>Maria Teresa Yuren Camarena</nombre>
                <institucion>Universidad Pedagogica Nacional</institucion></autor>
                <titulo>Educacion centrada en valores y dignidad humana</titulo>
        </portada>
        <texto>
                <div0><encabezado>Presentacion</encabezado>
                        <p>La imagen esperanzadora que perfilaron no pocos intelectuales,
                                politicos, banqueros y empresarios a partir del termino de la guerra fria,
                                del supuesto "<enfatizado>fin de las ideologías</enfatizado>" y de la
                                idea de la ""<enfatizado>aldea global"</enfatizado>", se desvanecio
                                bajo los trazos que la realidad fue delineando como signos
                                de una creciente tendencia a la barbarie que pone en peligro la vida en
                                el planeta y que lesiona, de muchas maneras y en muchos rincones de
                                la tierra, la dignidad humana. Ante este panorama, se ha vuelto
                                insistente la pregunta respecto de si los procesos educativos escolares
                                pueden contribuir en algo para oponerse a esa tendencia. La respuesta
                                positiva a esta interrogante se ha traducido en diversos esfuerzos para
                                desarrollar lo que se ha llamado "<enfatizado>educacion valoral
                                </enfatizado>", o "<enfatizado>educación centrada en valores
                                </enfatizado>"; esfuerzos que si bien difieren en el signo teorico e
                                ideologico y en el tipo de estrategia aplicada, han estimulado la
                                interlocucion, la reflexion y el estudio sobre el tema. </p>
                        <p>El presente trabajo tiene la intencion de contribuir a ese debate
                                teorico al aportar algunos elementos que sustentan la tesis de que una
                                educacion escolar centrada en valores ha de ser un proceso
                                eminentemente formativo que tenga como horizonte axiologico la
                                dignidad humana. </p>
                </div0>
…
                <notas>NOTAS
                        1 Baste nombrar: el deterioro ambiental persistente que se ve acelerado
                        por los accidentes nucleares y los desastres ecologicos; la pobreza
                        extrema de una amplia capa de la poblacion mundial que aumenta en
                        proporcion inversa a la hiperconcentracion de la riqueza; los racismos y
                        fundamentalismos que se han traducido en violencia brutal; el aumento
                        alarmante del desempleo,la delincuencia y la drogadiccion; las guerras
                        civiles prolongadas y las guerras de baja intensidad que asuelan diversas
                        regiones del planeta; la incapacidad de las instituciones politicas,
                        economicas y sociales para contribuir a la satisfaccion de las necesidades
                        materiales y espirituales de la humanidad, y las actitudes egoticas, la
                        venalidad, la banalidad y el consumismo que, atados al modelo
                        economico, se globalizan junto con este.
                        2 Agnes Heller. Teoria de las necesidades en Marx. Tr. J. Ivars.
                        Barcelona,
                        Peninsula, 1958. 182 p., p. 157 y ss.
…
                </notas>
                <bibliografia>Bibliografia
                        <itemBib>Habermas, Jurgen. Conciencia moral y accion comunicativa. Tr.
                                R. Garcia. Barcelona, Peninsula, 1985 (Col.Homo Sociologicus, num.
                                24), 219 p. </itemBib>
                                ….
                        <itemBib>Heller, Agnes. Teoria de las necesidades en Marx. Tr. J. Ivars.
                                Barcelona, Peninsula. 1958, 182 p., p. 157 y ss.
                                ….
                </bibliografia>
```

```
        </texto>
</articulo.1>
```

Figura 4.7. Artículo marcado con el vocabulario articulo.dtd

Una vez establecido el DTD y marcados algunos artículos se procedió a elaborar hojas de estilo XSD para verificar el despliegue en formato HTML vía un browser. En la figura 4.7 puede verse el XSD que se utilizó para las pruebas con usuarios.

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
        <xsl:template match="articulo.1">
                <HTML>
                        <HEAD>
                                <META content="text/html;charset=ISO-8859-1"/>
                                <TITLE>Revistas de Educacion y pedagogia </TITLE>
                        </HEAD>
                        <BODY>
                                 <H1 align="CENTER">
                    Revistas de Educacion
                 </H1>
                                <H2 align="center">
                                        <xsl:value-of select="relaciones"/>
                                </H2>
                                <H3 align="center">
                                        <xsl:value-of select="portada/titulo"/>
                                        <BR/><BR/>
                                        <xsl:value-of select="portada//nombre"/>
                                         <BR/>
                        <xsl:value-of select="portada//cargo"/>
                         <BR/>
                        <xsl:value-of select="portada//institucion"/>
                        <BR/>
                        <xsl:for-each select="portada//email">
                        <xsl:value-of select="."/>
                        <BR/>
                        </xsl:for-each>
                       <!--xsl:apply-templates select="portada//email"/-->
                                </H3>
                                <H4>
                               <xsl:for-each select="portada//resumen/p">
                    Resumen:
                    <xsl:value-of select="."/>
                </xsl:for-each>
                <BR/>
                <BR/>
                <xsl:for-each select="texto/p">
                    <xsl:value-of select="."/>
                    <BR/>
                </xsl:for-each>
                <xsl:for-each select="texto/div0">
                    <xsl:value-of select="."/>
                    <xsl:for-each select="texto/div0">
```

```
                        <xsl:value-of select="texto/div0/encabezado"/>
                        <BR/>
                    </xsl:for-each>
                    <xsl:for-each select="texto/div0/p">
                        <xsl:value-of select="."/>
                        <BR/>
                    </xsl:for-each>
                </xsl:for-each>
                <BR/>
                <xsl:for-each select="texto/bibliografia">
                    <xsl:value-of select="."/>
                    <BR/>
                </xsl:for-each>
                                </h4>
                                <HR />
                                Copyright 2003
                                <BR />
                                Alma Beatriz Rivera Aguilera, Universidad Iberoamericana
                                <BR />
                                <A HREF="mailto:alma.rivera@uia.mx"> alma.rivera@uia.mx</A>
                            </BODY>
                    </HTML>
            </xsl:template>
            <!--xsl:template match="portada//email">
                <xsl:value-of select="."/>
            </xsl:template -->
</xsl:stylesheet>
```

Figura 4.8 Hoja de estilo articulo.xsl referida al artículos marcados con el vocabulario articulo.dtd

Las figuras 4.6, 4.7 y 4.8 muestran los elementos más importantes de la implementación del vocabularios artículo: DTD, artículos marcados y hojas de estilo que convierten el texto en formato .xml a formato .html.

Fase 5: **Validar el resultado y verificar que se están alcanzando los objetivos**

En este punto se revisó la lista de elementos, el código del DTD y se marcaron 29 documentos. Durante el marcado saltaron algunos puntos sobre las etiquetas y el posterior indizado acerca de los cuales se tomaron las siguientes decisiones:

1. Con respecto a la nomenclatura de elementos se pueden tomar dos opciones: nombrarlos de forma que todo elemento esté asociado a un contexto o mantener los nombres independientes del contexto. En cualquiera de los dos casos el indizado y despliegue de los elementos

puede llevarse a cabo. En este vocabulario se decidió diferenciar ciertas etiquetas tales como **nombre** o **titulo**, ya que aparentemente son lo mismo pero el nombre de un autor de un artículo, para efectos de recuperación de contenidos, será diferente al nombre de uno inmerso en la narrativa del texto y otro citado en la bibliografía. Otras etiquetas se mantuvieron independientes del contexto como el caso de la etiqueta **lista**, **tabla** o **figura** pues ya sea anidadas en la etiqueta **texto** o en el nivel **texto/div0** hasta **div5** reflejan el mismo tipo de información. (Maler, E. y El Andaloussi, J., 1996, pp. 184-185). Con respecto al número de caracteres de un nombre de etiqueta para SGML (Idem, p. 248) estaba establecido en 8 caracteres, para XML ese límite puede ser rebasado.

2. Para efectos de esta prueba no se utilizarán atributos, ni entidades
3. Si se necesita indizar a los autores por frase o para generar un listado por apellido será necesario desagregar por apellido y nombre. En estas pruebas se mantendrá el nombre en una sóla etiqueta.

Existen utilerías en Internet que permiten validar la conformación adecuada de los archivos tanto DTD´s como los de tipo XML. Uno de estos sitios es el de la Universidad de Brown: http://www.stg.brown.edu/service/xmlvalid en donde se hicieron validaciones de los archivos. También se validaron tanto a nivel de archivos bien formados (“well-formedness”) como válidos (“validated”) de acuerdo a los editores que se utilizaron en el marcado de los archivos: XmlSpy y XMLWriter.

Fase 6. **Documentar el esquema y capacitar a la gente que lo utilizará**

La documentación del esquema se muestra en los anexos 1 y 2, diagrama de árbol del esquema y formas de elementos respectivamente; así como todo el apartado 4.1 de esta tesis en donde se ha descrito el diseño e implementación de este vocabulario.

Dadas las características de esta investigación no se desarrolló un programa de capacitación.

4.2 Colección de artículos de revista

El planteamiento inicial para generar la colección de artículos fue la de tomar algunos de los artículos de revista publicados en HTML en el sitio de la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES)[7] tomando una muestra de artículos de diferentes disciplinas y tomando como evaluadores a los investigadores registrados en el SNI de la comunidad académica de la UIA. Se consideró que la variedad de disciplinas requeriría una cantidad de artículos que no sería posible marcar en el período de un mes que se programó para ello. Dadas las limitaciones de recursos se decidió marcar exclusivamente revistas sobre el tema de educación ya que la universidad cuenta con un Instituto de Investigaciones para el Desarrollo de la Educación (INIDE), un Departamento de Educación y un área de Didáctica dentro del Centro de Formación Valoral; todas estas unidades académicas con personal experto en el tema de educación.

Se marcaron 28 artículos tomados de 3 revistas cuya temática abarca la educación, cada artículo es un archivo por separado:

[7] http://www.hemerodigital.unam.mx/ANUIES/

1. DIDAC de la Universidad Iberoamericana, un número, 8 artículos
2. La Academia del Instituto Politécnico Nacional , dos números, 9 artículos
3. Pedagogía de la Universidad Pedagógica Nacional, dos números, 11 artículos.

Los archivos ya existían originalmente en formato electrónico, pero no poseían marcas. Se llevaron a cabo las primeras pruebas utilizado el editor NotePad de Microsoft. Una vez se consolidó la DTD se necesitó un editor adecuado para XML que permitiera un marcado más sencillo, se buscó un editor de software libre y se seleccionó XMLSpy[8] el cual ofrece una versión de prueba por un mes y XMLWriter[9] fue la segunda prueba gratis que se utilizó.

Se identificaron ciertas características de los textos que tuvieron incidencia en el planteamiento del flujo de trabajo para el marcado.

- Algunos artículos erán un solo archivos y otros estaban separados en archivos diferentes para cada parte del artículo.
- Los archivos estaban en HTML y se convertirán a .TXT para marcarlos posteriormente con XML
- En La Academia se encontró un artículo copiado de otra publicación, sin los datos exactos de la fuente y con citas inmersas pero sin la bibliografía correspondiente

Durante el proceso de marcado surgió la inquietud del marcado automático, aunque sale del marco de esta investigación. Para algunos elementos como párrafo el marcado podría llevarse a cabo con un programa ad-hoc. El marcado

[8] http://www.xmlspy.com/
[9] http://xmlwriter.net/

inverso[10] también resulta interesante cuando un tema es indicado a nivel de metadatos y ese concepto se busca en el texto y se marca en ese contexto con una etiqueta como **tema** de este vocabulario.

El marcado automático del resto de los documentos dependería de si los archivos que se tomen como origen tienen una estructura uniforme; en cualquier caso las marcas tendrían que ser revisadas y complementadas por un individuo.

Como ya se indicó el principal interés en esta investigación es el marcado del llamado cuerpo del texto de los artículos de revista. Las dos restantes áreas son al inicio los metadatos descriptivos y al final la bibliografía. Se considera, como también se indicó con anterioridad, que los metadatos pueden registrarse en cualquier otro vocabulario adecuado, como el Dublín Core , el Front de SciELO o de JAI-JP, o el TeiHeader del TEI (Tabla 4.3). Con respecto a la bibliografía se puede considerar el TEI, aunque sea limitado, o las de SciELO que pueden ser demasiado extensas y engorrosas; el JAI es mas sensato para efectos de marcado de bibiografías, pero tiene elementos demasiado detallados en mi opinión como “first page” y “last page”, De cualquier manera la bibliografía y el front no son lo mas importante aquí, sino el contenido; por lo que el vocabulario y las marcas en los artículos, si bien se han indicado, no han sido extensos en lo referente a portada y bibliografía. En este sentido el vocabulario quedaría como un espacio de nombre en la parte estructural y un perfil de aplicación en el encabezado descriptivo y la parte bibliográfica.

[10] Idea expresada por la Mtra. Clara López en una conversación llevada a cabo durante las asesorías para esta tesis.

4.3 Instalación y configuración de Harvest

La instalación y configuración de Harvest se llevó a cabo con la ayuda del manual oficial (Hardy, D.R., et.al., 29 de octubre de 2002) y la amable asistencia de expertas en Harvest de la DGSCA-UNAM[11]. La instalación consistió en los siguientes pasos:

- Identificar cuál era la versión más confiable, la cual se nos indicó en la UNAM-DGSCA se trataba de la 1.8.2
- Se identificó un servidor de Linux con las características indicadas en el manual de Harvest y se solicitó una cuenta con los permisos necesarios para la instalación[12]
- Se bajó el software a través de la lista de distribución en el sitio http://prdownloads.sourceforge.net/harvest/
- Se descomprimió el código, se corrieron los comandos de instalación y se configuraron los archivos de Apache necesarios. En este punto fue necesario asegurarse que la cuenta de instalación tuviera los permisos necesarios para las operaciones de creación de directorios y modificación de archivos, para ello fue de gran ayuda contar con el apoyo del administrador del equipo.
- Se generó el directorio necesario para poder ver los archivos XML vía web

[11] Ingenieras Angélica Denise Corral y Rosa García, Jefe de Programación y Programadora de la Coordinación de Publicaciones Digitales de la UNAM-DGSCA.
[12] Se contó con el apoyo del Ing. Héctor Masayuki Hernández y del Mat. José Juan Téllez, del área de Automatización de la Biblioteca Francisco Xavier Clavigero de la UIA., para llevar a cabo estas tareas.

- Se crearon dos Gatherers y Brokers para generar los dos espacios de búsqueda de los archivos XML; tanto para índice de texto completo como para índice por ciertas etiquetas.
- En cada uno de los interfases se configuró la aplicación a través de los archivos necesario utilizando como referencia el Manual de Harvest de García, R.(2003).
- Se probaron las aplicaciones que se encuentran disponibles en las direcciones http://kirin.bib.uia.mx/Harvest/brokers/revistas/summary.html para la versión que busca sobre el texto completo y http://lyncis.dgsca.unam.mx/Harvest/brokers/alma/summary.html para la versión que busca por etiquetas. La segunda interfase se encuentra en el servidor de la UNAM-DGSCA debido a que la configuración por etiquetas no corrió adecuadamente en el servidor de la Biblioteca FXC de la UIA para las fechas en las que se llevaron a cabo las entrevistas. Las entrevistas se decidió llevarlas a cabo exclusivamente en el servidor de la UIA y sobre texto completo debido a que la evaluación de esos resultados era suficiente para inferir la evaluación por etiquetas y a que las estrategias de búsqueda por etiquetas son bastante engorrosas[13].

Las etiquetas que conformaron el índice de marcado interno fueron:

[13] Con mayor tiempo disponible se hubiese podido desarrollar un interfase más amigable para la búsqueda por etiquetas, pero dado que los objetivos de la investigación se cumplían con el interfase en texto completo, el tiempo disponible era escaso y no poseo “expertise” en programación de interfases web, decidí utilizar solo la aplicación en texto completo durante las pruebas.

autor	item
citaText	lugar
definición	pieFigura
ejemplo	resumen
encabezado	siglas
enfatizado	tema
fecha	titulo
institucion	

Tabla 4.6 Elementos a indizar.

4.4 Recolección de datos exploratorios sobre el impacto del marcado en la precisión.

En el apartado 3.3 Diseño experimental se describe en detalle el método de recolección de datos, en este apartado se reportarán los resultados de las entrevistas.

Se llevaron a cabo 8 entrevistas con sendos expertos en el tema de la educación, en la tabla 4.7 se describen algunas de sus características

Licenciados	Maestros	Doctores	+ 5 años de Exp.	+ 10 años de Exp.	+15 años de Exp.	+20 años de Exp.
3	3	2	3		1	4

Tabla 4.7 Características de los expertos en educación que evaluaron la precisión de los artículos recuperados

De los 3 licenciados 2 son pasantes de maestría y uno de los maestros es pasante de doctorado.

Las sesiones duraron entre medio hora y una hora, dependiendo del tiempo que cada experto necesita para interactuar con el sistema; y se llevaron a cabo en sus

oficinas. Cabe mencionar que en 3 de los casos los equipos no poseían browsers con la capacidad de desplegar archivos XML utilizando hojas de estilo XSL (Explorer y Netscape 5 o menos), en esos casos los investigadores ante la opción de cambiar de equipo o evaluar el documento en su formato xml optaron por lo segundo.

Se llevaron a cabo 3 búsquedas por cada evaluador, a excepción de un evaluador que sólo pudo completar 2, con lo que se dieron 23 estrategias con los correspondientes resultados.

El promedio de precisión para la búsqueda de texto libre fue de 40.72 % lo que representa de que por cada 10 documentos recuperados 4 documentos podrían ser útiles para el investigador.

ID de la búsqueda	Resultados Texto Libre	Relevantes	Precisión Texto Libre	Resultados Etiquetas	Relevantes	Precisión Etiquetas
L-1	12	2	16.67%	1	1	100.00%
L-2	14	5	35.71%	2	0	0.00%
L-3	4	2	50.00%	1	1	100.00%
G-1	14	8	57.14%	3	2	66.67%
G-2	5	0	0.00%	2	0	0.00%
G-3	5	4	80.00%	3	3	100.00%
P-1	11	5	45.45%	3	1	33.33%
P-2	4	2	50.00%	0	0	
P.3	14	4	28.57%	3	1	33.33%
M-1	8	1	12.50%	1	0	0.00%
M-2	4	0	0.00%	0	0	
M-3	4	0	0.00%	0	0	
V-1	2	2	100.00%	1	1	100.00%
V-2	14	7	50.00%	6	2	33.33%
V-3	8	2	25.00%	1	1	100.00%
F-1	5	1	20.00%	0	0	
F-2	8	5	62.50%	4	3	75.00%
F-3	5	0	0.00%	0	0	
E-1	12	8	66.67%	3	3	100.00%
E-2	1	1	100.00%	0	0	
E-3	7	5	71.43%	2	2	100.00%
A-1	14	4	28.57%	6	4	66.67%
A-2	11	4	36.36%	5	3	60.00%
Promedio			40.72%			62.84%
Desv.Stand.			4.62			3.88

Tabla 4.8 Mediciones de precisión totales. Cuando el resultado de la recuperación fue de 0 hits se eliminó la búsqueda del promedio. Para texto completo las búsquedas fueron 23 y para etiquetas las búsquedas consideradas fueron 17.

Como puede observarse en la tabla 4.8 los resultados en la precisión al utilizar las marcas del vocabulario articulo suben hasta el 62.84%, eso indica que de cada 10 potenciales artículos recuperados 6 son evaluados como útiles. Al eliminar los casos en que no es posible medir la precisión de la búsqueda con etiquetas (Tabla 4.9) se tiene que para 17 casos la precisión de la búsqueda de texto libre es de 45.09% y la de etiquetas naturalmente se mantiene en 62.84, con ello tenemos

que aunque la precisión para búsqueda en texto libre se incrementa de 40.72% a 45.09% sigue siendo mayor la precisión con etiquetas y es de 62.84%.

Cabe recordar que los datos de este reporte son exploratorios y sólo indican una tendencia, por lo que no pueden considerarse como probatorios.

ID	Result. Texto libre	Relevantes	Precisión textolibre	Result. Etiquetas	Relevantes	Precisión Etiquetas
L-1	12	2	16.67%	1	1	100.00%
L-2	14	5	35.71%	2	0	0.00%
L-3	4	2	50.00%	1	1	100.00%
G-1	14	8	57.14%	3	2	66.67%
G-2	5	0	0.00%	2	0	0.00%
G-3	5	4	80.00%	3	3	100.00%
P-1	11	5	45.45%	3	1	33.33%
P.3	14	4	28.57%	3	1	33.33%
M-1	8	1	12.50%	1	0	0.00%
V-1	2	2	100.00%	1	1	100.00%
V-2	14	7	50.00%	6	2	33.33%
V-3	8	2	25.00%	1	1	100.00%
F-2	8	5	62.50%	4	3	75.00%
E-1	12	8	66.67%	3	3	100.00%
E-3	7	5	71.43%	2	2	100.00%
A-1	14	4	28.57%	6	4	66.67%
A-2	11	4	36.36%	5	3	60.00%
Promedios			45.09%			62.84%
Desv.Stand.			3.91			3.88

Tabla 4.9 Mediciones de precisión eliminando los búsquedas que tuvieron 0 hits para etiquetas. Tanto para texto completo como para etiquetas las búsquedas consideradas fueron 17.

Los datos arrojaron información para calcular otras dos medidas interesantes:

- Razón de recuperación o recall, establecida entre los aciertos totales en la búsqueda en texto completo y los aciertos en la búsqueda con etiquetas. Está razón fue en promedio de 36.11 % lo cual indica que sólo la tercera parte de los artículos útiles, en promedio, fue recuperada con etiquetas.

ID	Aciertos texto completo	Aciertos etiquetas	Razón de recuperación	Falsos aciertos texto completo	Falsos aciertos etiquetas	Razón de falsos aciertos recuperados
L-1	2	1	50.00%	10	0	0.00%
L-2	5	0	0.00%	9	2	22.22%
L-3	2	1	50.00%	2	0	0.00%
G-1	8	2	25.00%	6	1	16.67%
G-2	0	0	0.00%	5	2	40.00%
G-3	4	3	75.00%	1	0	0.00%
P-1	5	1	20.00%	6	2	33.33%
P-2	2	0	0.00%	2	0	0.00%
P.3	4	1	25.00%	10	2	20.00%
M-1	1	0	0.00%	7	0	0.00%
M-2	0	0		4	0	0.00%
M-3	0	0		4	0	0.00%
V-1	2	1	50.00%	0	0	
V-2	7	2	28.57%	7	4	57.14%
V-3	2	1	50.00%	6	0	0.00%
F-1	1	0	0.00%	3	0	0.00%
F-2	5	3	60.00%	3	1	33.33%
F-3	0	0		5	0	0.00%
E-1	8	3	37.50%	5	0	0.00%
E-2	1	0	0.00%	0	0	
E-3	5	2	40.00%	2	0	0.00%
A-1	4	4	100.00%	11	2	18.18%
A-2	4	3	75.00%	7	2	28.57%

Promedios		36.11%		12.83%
Desv.Stand.		4.08		4.45

Tabla 4.10 Razón de recuperación (recall) de los hits producto de las búsquedas con etiquetas y los hits de las búsquedas en texto completo. Cuando el resultado de los aciertos y falsos aciertos fue 0 se eliminó la búsqueda del promedio. Para razón de recuperación las búsquedas fueron 20 y para la razón de falsos aciertos las búsquedas fueron 21.

- Razón de falsos aciertos recuperados, calculada como el porcentaje de falsos aciertos, evaluados como tales en la búsqueda en texto completo, y que vuelven a surgir como resultado de la búsqueda con etiquetas. Esta razón resultó en promedio con un valor de 12.83, lo cual implica que un de cada 10 falsos aciertos recuperado con texto completo sólo uno, en promedio, aparece en la búsqueda de etiquetas.

En las tablas 4.9 y 4.10 puede apreciarse que también se calcularon las desviaciones estándar, las cuales indican los puntos hacia arriba y hacia abajo en que varían los promedios que en este caso varía desde 3.88 a 4.62 en el caso de la medición de precisión.

5. CONCLUSIONES Y RECOMENDACIONES

En los capítulos anteriores se recorrió una ruta que va desde el problema de la recuperación de información hasta los resultados a nivel de precisión que puede ofrecer un lenguaje de marcado. El marco de la investigación se registró en el capítulo 2, y el capítulo 3 estableció los elementos metodológicos de la investigación. El capítulo cuatro incluyó los aportes fundamentales de esta tesis: el diseño e implementación de un vocabulario XML para el marcado de artículos de revista, la configuración del buscador Harvest y las pruebas con usuarios para evaluar la precisión de los resultados de búsqueda sobre archivos marcados, tanto en su modalidad de todo el texto o sólo sobre determinadas marcas.

En este capítulo de conclusiones se regresa a los planteamientos iniciales; y se contrastan resultados descritos en el capítulo 4 con la hipótesis y objetivos propuestos en los capítulos 1 y 2.

5. 1 Conclusiones con relación al problema de investigación, los objetivos y la hipótesis.

El problema de investigación fue planteado a partir de la dificultad real de recuperar información de forma precisa por los usuarios en grandes repositorios de documentos en texto completo; y consistió en:

> *Definir un vocabulario XML que describa adecuadamente la estructura de artículos de revistas de tipo académico en bibliotecas digitales y recoja los contenidos significativos inmersos en el texto; de forma tal que al aprovechar este marcado a través de un buscador la precisión en los resultados de una búsqueda sea más alta que si el buscador indiza la totalidad del texto*

El problema expuesto fue atacado a través del cumplimiento de los objetivos abajo descritos. La definición del vocabulario XML para artículos de revista enfocado a elementos estructurales y de contenido inmersos en el texto es un aporte al área de aplicaciones XML en publicación digital, ya que no existía al inicio de la investigación un vocabulario con esas características y en idioma español.

El objetivo principal fue:

> *Proponer y evaluar un vocabulario XML que describa y estructure adecuadamente los contenidos de artículos de revistas académicas, dicha descripción y estructuración potenciará los archivos con las bondades del XML, en especial se espera mejore la recuperación de información en términos de incremento en la precisión de los resultados. Un objetivo auxiliar es probar una metodología adecuada para el diseño de dicho vocabulario, la de Maler, E. y El Andaloussi, J., (1996) para el desarrollo de DTDs o esquemas.*

Con respecto a los objetivos se consideran cumplidos, tal y como se describió en el capítulo 4 se desarrolló un vocabulario para artículos de revista, utilizando la metodología de Maler y El Andaloussi. La evaluación se llevó a cabo con la salvedad de que se trata de un estudio exploratorio, tal y como se planteo en la metodología establecida en el capítulo 3.

Un aporte al campo del desarrollo de aplicaciones de lenguaje de marcado es el ejercicio de la aplicación de la metodología de Maler y El Andaloussi y la práctica de desarrollo de un diseño de forma local, ya que existe una enorme tendencia a adaptar vocabularios ya desarrollados y probados en otros países. Ciertamente es importante antes de llevar a cabo un desarrollo, investigar y conocer las

aplicaciones ya existentes; pero en este caso, después de una revisión exhaustiva de vocabularios ya diseñados, se llegó a la conclusión de un desarrollo propio era necesarios, para comprobar la hipótesis propuesta. Este avance puede romper brecha al demostrar que pueden llevarse a cabo diseños e implementaciones locales y ojalá comunitarias en el futuro.

La hipótesis principal de la investigación fue:

> *Al marcar artículos de revista con el vocabulario XML adecuado, los resultados de la búsqueda que aproveche el indizado sobre ciertas partes marcadas del texto de los artículos, serán más precisos que la misma búsqueda sobre el texto completo.*

Esta hipótesis fue encontrada verdadera en el contexto del estudio exploratorio ya que al marcar un grupo de documentos con el vocabulario y llevar a cabo búsquedas tanto de texto completo, como sobre las etiquetas marcadas se tuvo como resultado que la medición de precisión fue mayor en promedio para las búsquedas sobre etiquetas, tanto en las mediciones sobre todos los casos (23) como en el promedio calculado eliminando los casos de 0 aciertos (17).

	Sin casos 0 aciertos		Todos los casos	
	Precisión textolibre	Precisión Etiquetas	Precisión textolibre	Precisión Etiquetas
Promedios	45.09%	62.84%	40.72%	62.84%
Desv.Stand.	3.91	3.88	4.62	3.88

Tabla 5.1 Resultados de precisión.

Lo anterior representa una tendencia hacia la comprobación de la hipótesis, pero de ninguna manera puede tomarse como un resultado probatorio. Esta tendencia es coherente con el estudio INEX (the INitative for the Evaluation of XML retrieval) (Govert, N. y Kazai, G. ,2003, p. 15) en donde los resultados en la precisión de

diferentes instituciones participantes en la evaluación mundial de buscadores aprovechando artículos marcados con XML es siempre más alta al aprovechar el marcado con datos que van desde el 15 al 34 % para XML y del 5 al 9% para el texto completo.

5.2 Implicaciones para la teoría.

Como se registró ampliamente en el capítulo 2, diversos autores (Witten, I.H. y Bainbridge, D. 2003, cap. 5; Harold, E. R. 2001: Duke, J.K. 1989; Boiko, B. 2001b) han considerado el impacto positivo de los lenguajes de marcado en la mejora de los resultados de búsqueda, pero pocas investigaciones reportan datos concretos al respecto, el estudio INEX (Govert, N. y Kazai, G.,Idem) es uno de los más importantes en esa línea. El enfoque de esta investigación fue el uso del lenguaje de marcado en la estructura del documento y en ciertas palabras o grupos de palabras que pudiésen reflejar el contenido del mismo. La combinación estructura-contenido en el marcado es novedosa y al mostrar los resultados del estudio una tendencia, a la comprobación de la hipótesis de una relación positiva entre aprovechamiento de lenguaje de marcado por un buscador y resultados de precisión, se considera un aporte interesante y un incentivo para estudios posteriores de carácter probatorio.

5.3 Implicaciones para políticas y prácticas.

Esta investigación tuvo que acotarse para poder llegar a un resultado útil en un tiempo limitado a un año de un investigador. Durante el trabajo surgieron gran cantidad de temas y facetas de interés para la discusión, implementación y el trabajo cooperativo. Es posible que algunas de estas inquietudes ya estén siendo

atacadas por individuos e instituciones, peor no se tiene noticia de ello. Vale la pena mencionar que es necesario en el ámbito de políticas y prácticas:

- Dar a conocer a los actores involucrados en la edición y uso de revistas electrónicas este vocabulario y proponer su discusión y mejora. En caso de llegar a un consenso registrar un espacio de nombre.
- Establecer el uso de algún estándar referentes a marcado estructural para publicaciones electrónicas.
- Desarrollar estándares para interoperabilidad similares a OAI pero basado en metadatos estructurales, o proponer el enriquecimiento de OAI con metadatos estructurales.
- Aplicación de marcado estructural y buscadores a documentos generalizados o multimedia.
- Desarrollo de un equivalente a RDF: RSF Resource Structural Format?
- Colaboración intrainsititucional e interinstitucional para el desarrollo de vocabularios a nivel de comunidades de usuarios y de investigaciones sobre la relación entre la recuperación de información y el marcado estructural.
- Llevar a cabo ejercicios de diseño e implementación como el aquí descrito, por equipos en los que participen tomadores de decisiones relacionados con la publicación y difusión de publicaciones electrónica en general y revistas académicas en particular con el fin de ir conociendo los vocabularios existentes y desarrollando nuevos estándares.

5.4 Limitaciones.

Este estudio por ser de carácter exploratorio no puede considerarse un aporte definitivo, sino más bien la identificación de una tendencia como se reportó en el apartado 5.2. En el apartado 3.2 se detallan las características metodológicas que este estudio hubiese requerido para tener carácter probatorio, entre ellas el establecimiento de una muestra probabilística de los artículos de revista de tipo académico que formen parte de bibliotecas digitales en México. Dadas las limitaciones de tiempo y presupuestales se optó por utilizar un pequeño grupo de 29 artículos de revistas de educación y se evaluó la precisión a través de 8 académicos especialistas en educación de la UIA.

El perfil de la responsable de esta tesis no incluye la programación de interfases en web, lo cual limitó la adecuación amigable de los interfases de búsqueda de Harvest, pero no menoscabó el proceso de medición de pertinencia. Cabe considerar que aun para un experto en programación Java este tipo de desarrollos hubiese tomado un tiempo considerable.

5.3 Investigaciones posteriores

Esta investigación empuja una línea de problemas de investigación en relación con los lenguajes de marcado y la recuperación de información, por su mismo carácter exploratorio es necesario continuar investigando con muestras más significativas y diseños consensuados en equipos interdisciplinarios. Además de esa vertiente principal a futuro surgen inquietudes académicas complementarias como las siguientes:

- Estudiar la eficiencia de los sistemas de recuperación de información textual en relación al marcado estructural.

- Comparar de tecnología de bases de datos, buscadores tradicionales de texto y buscadores específicos sobre archivos marcados con XML
- Profundizar de las posibilidades de Xlink y XPath en la generación y aprovechamiento de los vocabularios XML, lo cual no se ha explotado en el vocabulario aquí propuesto.
- Evaluar buscadores que aprovechen XPath y XQuery en relación al vocabulario propuesto.
- Llevar a cabo una investigación con medidas de relevancia gradual o escalonada: no relevante, algo relevante, relevante, muy relevante, perfectamente relevante.
- Evaluar a la brevedad el uso del TEI y sus versiones multilingües para considerar el aprovechamiento de esa tecnología de marcado y los desarrollos alrededor del mismo.
- Considerar a partir del vocabulario propuesto en esta investigación para artículos de revista, la posibilidad de desarrollar un vocabulario genérico para textos académicos comunes en bibliotecas digitales universitarias.
- Considerar las búsquedas como frase en pruebas futuras ya que no fueron integradas a las estrategias de búsqueda que aquí se llevaron a cabo.
- Llevar a cabo pruebas de interoperabilidad entre repositorios de archivos marcados con este vocabulario y albergados en diferentes servidores.
- Búsqueda de alternativas en la medición de la relevancia, que pueden basarse más en productos que en apreciaciones subjetivas. Por ejemplo,

uso de los resultados de búsqueda en bibliografías de trabajos, tesis, publicaciones, etc. A fin de darle un aspecto menos subjetivo al concepto

BIBLIOGRAFÍA

Ahmed, K. et. al. (2001). *XML Meta Data.* Birmingham, UK: Wrox Press.

Arms, W.Y. (2000). *Digital libraries.* Cambridge, Massachusetts: MIT Press.

Baeza-Yates, R., Fuhr, N. Y Maarek, Y.S. (2002*). Second Edition of the "XML and Information Retrieval" workshop.* Llevado a cabo durante SIGIR 2002, Tampere, Finlandia, Agosto 15 2002. Recuperado de: http://www.acm.org/sigs/sigir/forum/F2002/maarek.pdf el 5 de junio de 2003

Barker, J. (2002). *Beginning Java objects*. Birmingham, UK: Wrox Press.

Beaubien, R. (2001). *Making of America II Project*. Berkeley: University of California Berkeley. Recuperado de: http://sunsite.berkeley.edu/moa2/ el 15 de enero de 2003.

Belew, R. K (2000). *Finding out about: a cognitive perspective on search engine technology and the WWW*. Cambridge : Cambridge University Press.

Bia, A.G., Sánchez-Quero, M. y Deau, R. (2003, mayo). *Multilingual Markup of Digital Library Texts Using XML, TEI and XSLT.* Ponencia presentada en la Conferencia XML Europa 2003, Londres, Inglaterra. Recuperado el resumen de: http://www.xmleurope.com/2003/wednesday.asp el 5 de junio de 2003.

Boiko, B. (2001b). *Content management bible*. New York: Hungry Minds.

Borgman, C.L., Moghdam, D. y Corbett, P.K. (1984). *Effective online searching*. New York : Marcel Dekker.

Borgman, C. L. (2000). *From Gutenberg to the global information infrastructure : access to information in the networked world*. Cambridge: MIT Press.

Buttenfield, B. (1999). Usability evaluation of digital libraries. En Stern, D. (1999). *Digital libraries : philosophies, technical design considerations, and example escenarios*. (pp. 39-59). New York: Haworth Press.

Burnard, L. y Sperberg-McQueen, C.M. (1995). *TEI Lite: Una introducción al Text Encoding for Interchange*. Documento No. TEI U5. Traducido por Manuel Sánchez Quero, Biblioteca Virtual Miguel de Cervantes, Universidad de Alicante, España. Septiembre 2001.

Carpenter, M. y Svenonius, E., (Eds.). (1985). *The conceptual foundation of descriptive cataloging*. San Diego: Academic Press.

Chapman, S. (2000). Considerations for project management. En Sitts, M. *Handbook for digital projects : a management tool for preservation and access*. (pp. 21-34). Andover, Massachussetts: Northeast Document Conservation Center.

Constantinopoulos, P. y Solverg, T. (Eds.). (2001). Research and advance technology for digital libraries. *Proceedings of the 5th European Conference on Digital Libraries en Darmstadt, Alemania*. Berlin: Springer-Verlag.

Cutter, C. A. (1985). Rules for a dictionary catalog : selections [Originalmente publicado en 1904 en Washington D.C por el Government Printing Office]. En M. Carpenter, y E. Svenonius, (Eds.). *The conceptual foundation of descriptive cataloging*. (pp. 62-71). San Diego: Academic Press.

Duke, J.K. (1989*).* Access and automation : the catalog record in the age of automation. En E. Svenonius, y M. Carpenter, (Eds.) *The conceptual foundation of descriptive cataloging*. (pp. 117-128) San Diego: Academic Press.

Elmasri, R. y Navathe, S.B. (2000). *Fundamentals of database systems*. 3a Ed. Reading, Massachusetts: Addison Wesley.

Fietzer, W. (2002). Interpretive encoding of electronic texts using TEI Lite. En W. Jones, J. R. Anronheim y J. Crawford. (Eds.) *Cataloging the web : metadata, AACR and MARC 21*. (pp. 103-108). Lanham, Maryland: Scarecrow Press / American Library Association.

Fox, E. y Urs, S.R. (2002). Digital libraries. En B. Cronin (Ed.) *Annual Review of Information Science and Technology* (pp. 503-590). Silver Spring, Maryland. American Society for Information Science and Technology.

Fuhr, N. et. al. (2001). Digital libraries: a generic classification and evaluation scheme. En P. Constantinopoulos y T. Solverg (Eds.). Research and advance technology for digital libraries, *Proceedings of the 5th European Conference on Digital Libraries in Darmstadt, Germany*. (pp. 187-199). Berlin: Springer-Verlag.

García, R. (2003). Manual de Harvest. Coordinación de Pubicaciones Digitales, Área de Programación, UNAM-DGSCA.

Gaynor, E. (1996). *From MARC to Markup: SGML and online library systems*.

Govert, N. y Kazai, G. (2003) *Overview of the Initative for the Evaluation of XML retrieval (INEX) 2002* Recuperado de: http://www.dcs.qmul.ac.uk/~gabs/papers/inex2002_overview.pdf el 5 de junio de 2003. El reporte con todos las aportaciones del seminario INEX 2002 puede verse en http://www.ercim.org/publication/ws-proceedings/INEX2002.pdf

Hardy, D.R., et.al. (29 de octubre de 2002). Harvest user´s manual. Recuperado de : http://harvest.sourceforge.net/harvest/doc/html/manual.html el 29 de agosto de 2003.

Harold, E.R. (2001). *XML Bible* (2a. ed.). New York : Hungry Minds.

Hernández Sampieri, R., Fernández Collado, C y Baptista Lucio, P. (1998). *Metodología de Investigación*. México: McGraw-Hill.

Jeliffe, R. (1998). *The XML and SGML cookbook : recipies for structured information*. Upper Saddle River, New Jersey: Prentice Hall.

Kenney, A.R. y Rieger O.Y. (1999). *Moving theory into practice : digital imaging for libraries and archive*s. Mountain View, California: Reserch Library Group.

Kieslig, K. (2002). Archival finding aids as metadata : Encoded Archival Description. En W. Jones, J. R. Anronheim y J. Crawford. (Eds.) *Cataloging the web : metadata, AACR and MARC 21*. (pp. 65-70). Lanham, Maryland: Scarecrow Press / American Library Association.

Lafuente López, R. Y Garduño Vera, R. (2001). *Lenguajes de marcado de documentos digitales de carácter bibliográfico*. Universidad Autónoma de México, Centro Universitario de Investigaciones Bibliotecológicas.

Lancaster, F.W. (1979). *Information retrieval systems: characteristics, testing and evaluation*. New York: Wiley.

Lancaster, F.W. (1983). *Evaluación y medición de los servicios bibliotecarios*. Universidad Autónoma de México, Dirección General de Bibliotecas. (Trabajo original publicado en 1977)

Lancaster, F.W. (1993) *If you want to evaluate your library*. University of Illinois.

Lassila, O. et. al. (1999). *Resource Description Framework (RDF) model and syntax specification*. W3C. Recuperado de http://www.w3.org/TR/REC-rdf-syntax/ el 19 de febrero de 2001.

Lesk, M (1997). *Practical digital libraries: books, bytes and bucks.* San Francisco: Morgan Kauffmann.

Li, C. y Stone, H.S (1999, enero). Digital library using next generation Internet. *IEEE Communications Magazine*. 70-71.

Li, G. y Huang, M.B. (2001). Research and development of digital libraries in Chine : majors issues and trends. En P. Constantinopoulos y T. Solverg (Eds.). Research and advance technology for digital libraries, *Proceedings of the 5th*

European Conference on Digital Libraries in Darmstadt, Germany. (pp. 450-457) Berlin: Springer-Verlag.

Luk, R.W.P et. al. (2002). A survey in indexing and searching XML documents. *Journal of the American Society for Information Science and Technology,* 53(6):415-437.

Maler, E. y El Andaloussi, J. (1996). *Developing SGML DTD´s: from text to model to markup*. Upper Saddle River, New Jersey: Prentice Hall.

Malo Alvarez, S. y Fortes Besprovani, M. (1999*). México frente a la era de la información*. Academia Mexicana de Ciencias.

Marcondes, C.H. y Sayao, L.F. (marzo, 2003). The ScieELO Brazilian scientific journal gateway and open archives. *D-Lib Magazine*. 9(3). Recuperado de http://www.dlib.org/dlib/march03/marcondes/03marcondes.html el 19 de marzo de 2003.

Marko, L. (2002). Working to a standard TEI headers for libraries. En W. Jones, J. R. Anronheim y J. Crawford. (Eds.) *Cataloging the web : metadata, AACR and MARC 21*. (pp. 53-55). Lanham, Maryland: Scarecrow Press / American Library Association.

Miller, E. y Hillmann, D. (2002). Libraries and the future of semantic web : RDF, XML, and the Alphabet Soup. En W. Jones, J. R. Anronheim y J. Crawford. (Eds.) *Cataloging the web : metadata, AACR and MARC 21*. (pp. 57-64). Lanham, Maryland: Scarecrow Press / American Library Association.

Morris, S. (2002). Comenzando la publicación de revistas electrónicas. En *Electronic Journal Publishing: A Reader 2.0*. INASP. Recuperado el 23 de abril de 2003 de http://www.inasp.info/psi/ejp/morrissp1.html

Morrison, M. et. al. (2000). *XML al descubierto*. Madrid: Prentice Hall.

Muller, M. (s.f.) *A very gentle introduction to the TEI*. Recuperado el 15 de marzo de 2003 de: http://faculty-web.at.nwy.edu/english/mmuller/ariadne/teixintro/teiman.htm

Rivera Aguilera, A.B. (2001). *Metadatos: viejos conceptos, nuevas realidades a la espera de unificar nomenclaturas y aplicar estándars*. Interfases 2001, Universidad de Colima. Colima, México.

Sánchez, A. y Fernández, M.L.(2000). Bibliotecas digitales en México*. Boletín de la Sociedad Mexicana de Ciencia de la Computación*. 1(2):13-16.

Sánchez Huitrón, J.A. (Julio-Diciembre 2002). Colecciones digitales universitarias en México. *Biblioteca Universitaria (Nueva Época).* 5(2): 130-143.

Schmierer, H.F.(1989). The impact of technology in cataloging rules. En E. Svenonius, y M. Carpenter, (Eds.) *The conceptual foundation of descriptive cataloging*. (pp. 101-116). San Diego: Academic Press.

Sitts, M.K. (2000). *Handbook for digital projects : a management tool for preservation and access*. Andover, Massachusetts: Northeast Document Conservation Center.

Stern, D. (1999). New search and navigation techniques in the digital library. En D. Stern (Ed.). *Digital libraries : philosophies, technical design considerations, and example escenarios*. (pp. 61-80). Haworth Press, New York.

Svenonius, E. (Ed.). (1985). *Foundations of cataloging : a source book*. Colorado: Libraries Unlimited.

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, Massachusetts: MIT Press.

Swan, A. Y Brown, S. (2002). *Authors and electronic publishing*. ALPSP. Citado por Morris, S. (2002). Comenzando la publicación de revistas electrónicas. En *Electronic Journal Publishing: A Reader 2.0*. INASP

Witten, I.H, Moffat, A. y Bell, T.C. (1999). *Managing gigabytes : compressing and indexing documents and images*. 2a. ed. San Francisco: Morgan Kaufman.

Witten, I.H. y Bainbridge, D. (2003). *How to build a digital library*. Amsterdam, Morgan Kaufman.

Wyke, R.A. y Watt, A. (2002). *XML Schema Essentials*. Nueva York: Wiley.

BIBLIOGRAFÍA COMPLEMENTARIA

Baldwin, C y Pullinger, D. (2000). What readers value en academic journals. Learned Publishing. 13(4):229-240. citado por Morris, S. (2002). Comenzando la publicación de revistas electrónicas. En *Electronic Journal Publishing: A Reader 2.0*. INASP

Beagle, D. (1999, Diciembre). Visualization of metadata*. Information Technology and Libraries*. 192-199.

Beaven Remnek, M. (1999). TEI analytical encoding and its application in the college classroom : the role of the library. *Technicalities.* 19(6), 9-12.

Beckett, D. et.al (2000). *An XML encoding of simple Dublin Core metadata*. Dublin Core Metadata Initiative. Recuperado de: http://dublincore.org/documents/2000/11/dcmes-xml/

Bia, A. G. y Carrasco, R. (2001). *Design and exploitation of a markup strategy in a digital library*. Documento no publicado. Recuperado de: http://cervantesvirtual.com/research/articles/pkdd2000.pdf el 5 de junio de 2003

Bishoff, L. y Garrison, W.A.(2000*). Metadata, cataloging, digitization and retrieval: who´s doing what to whom: the Colorado digitization project experience. For the Library of Congress bicentennial conference on bibliographic control for the new millennium: confronting the challenges of networked resources and the web*. [Washington: Library of Congress]. Recuperado de: http://lcweb.loc.gov/catdir/bibcontrol/bishoff_paper.html

Boiko, B. (2001a, octubre/noviembre). Understanding content management. *Bulletin of the American Society for Information Science and Technology*. 8-13.

Burnard, L. (2000, junio). Text Encoding for Interchange: a new Consortium. *Ariadne*, número 24. Recuperado de: http://www.ariadne.ac.uk/issue24/tei/intro.html el 4 de junio de 2003

Caplan, P. (2003). *Metadata fundamentals for all librarians*. Chicago. American Library Associations.

Carmel, D., Maarek, Y. y Soffer, A. *XML and information retrieval: a SIGIR 2000 workshop.* Recuperado de: http://www.acm.org/sigs/sigir/forum/S2000/XML_report.pdf el 5 de junio de 2003.

Chesnut, D.R. (1998, mayo). SGML and the digital libraries of tomorrow. *The Journal of Academic Librarianship*, 232-236.

Chudnov, D. (1999). Toward seamlessness with XML. In D. Stern (Ed.). *Digital libraries : philosophies, technical design considerations, and example escenarios*. (pp. 121-130). New York : Haworth Press.

Coombs, J.H., Renear, A.H., DeRose, S. *Markup Systems and the Future of Scholarly Text Processing*. Recuperado de http://www.oasis-open.org/cover/coombs.html el 5 de Junio de 2003

Doods, D. et. al. (2001). *Professional XML meta data*. Wrox Press.

Dowler, L. (1997*). Gateways to knowledge : the role of academic libraries in teaching, learning and research*. Cambridge, Massachussetts: MIT Press.

Dublin Core Metadata Initiative (1999*). Dublin Core Metadata Element Set, Version 1.1: Reference Description*. Recuperado de http://dublincore.org/documents/1999/07/02/dces/

Dublin Core Metadata Initiative (2000*). Dublin Core Qualifiers*. Recuperado de http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

Eman, J. ven (octubre/noviembre 2002). What can you do with XML today. *Bulletin of the American Society for Information Science and Technology*. 29 (1).

Recuperado de: http://www.lib.virginia.edu/speccol/scdc/articles/alcts_brieb.html

GILS: About a powerful, new way to find information (s.f.). Recuperado de: http://www.gils.net/about.html el 22 de enero de 2003.

Green, A., Dione, J. y Dennis, M. (1999*). Preserving the whole : a two track approach to rescuing social science data and metadata*. Washington: The Digital Library Federation. Recuperado de: http://www.clir.org/pubs/reports/pub83/pub83.pdf el 5 de julio de 2002.

Greenstein, D. (Nov/Dec 2001). *On Digital Library Standards: From Yours and Mine to Ours*. CLIR (Council of Library and Information Resources) Issues. 24. Recuperado de: http://www.clir.org/pubs/issues/issues24.html#digital

Heaps, H.S. (1978). *Information retrieval: computational and theoretical aspects*. Academic Press, New York.

Hillmann, D. (2001). *Using Dublin Core*. Recuperado de: http://dublincore.org/documents/2001/04/12/usageguide/

Jacsó, P. (2002, septiembre). XML and digital librarians. *Computers in Libraries*. Septiembre: 46-49.

Kim, Hun-Hee y Choi, Chang-Seok. (2000). XML: How it will be applied to digital library systems. *The Electronic Library*. 18(3), 183-189.

The Library of Congress. (mayo 9, 2003). *METS: an overview & tutorial*. Recuperada de http://www.loc.gov/standards/mets/METSOverview.html el 10 de junio de 2003.

López Guzmán, C. (2000). *Modelo para el desarrollo de bibliotecas digitales especializadas*. (Tesis de Maestría no publicada, Instituto Tecnológico Autónomo de México, 2000). Recuperado de: http://www.bibliodgsca.unam.mx/tesis/tes7cllg/tes7cllg.htm

Mallén, G (2000). New hashing algorithms for digital libraries. [CD ROM].*Proceedings of BitWorld 2000 Conference*. México: Universidad Iberoamericana.

Margulius, D.L.(2002). XML everywhere: as plataforms vendors incorporate XML, will content management problems go away? *Infoworld*. 10/28/02: 44-45.

Medeiros, N. (2000, septiembre/octubre). XML and the resource description framework : the great web hope. *Online*. 37-40.

Miller, D.R (2000, verano). XML: Libraries´ strategic opportunity*. Library Journal*. [Suplemento Net Connect], 18-22.

MOA2 Digital object document type definition tutorial. (s.f.). Recuperado de: http://sunsite.berkeley.edu/moa2/papers/dtdtutorial2.htm

Morgan, E.L.(2001). *Comparing Open Source Indexers* Recuperado el 28 de agosto del 2003 de : http://www.infomotions.com/musings/opensource-indexers/index.html

Rivera Aguilera, A.B. (2003). Administración de contenidos y XML en las colecciones digitales*. XXXIII Jornadas de Bibliotecología AMBAC 2003, Puerto Vallarta, México*. Asociación Mexicana de Bibliotecarios, A.C. Manuscrito presentado para su publicación.

Rogers, M. (2002, otoño). LC gets a new content management system. *Library Journal*. 6-7.

Saarela, J. J. (2000). *The role of metadata in electronic publishing*.[Resumen] Dissertation Abstracts International, 61(03). Tesis para el grado de Doctor en Tecnología de la Teknillinen Korkeakoulu, Finlandia.

Scielo Metodología [2001]. Recuperado de: http://www.scielo.org/dtd/ el 22 de enero de 2003

Simple guide for TEI lite XML markup. Recuperado de http://www.etext.leeds.ac.uk/cocoon/epb/lect/tei.xml el 5 de Junio de 2003.

Tennant, R. (2001, marzo). XML: The digital library hammer. *Library Journal*. 30-32.

Tennant, R. (2001, julio). The $64,000 question. *Library Journal*. 34.

The Commission on Preservation and Access y Research Libraries Group. (1996). *Preserving digital information : Report of the task force on archiving of digital information*. Recuperado de: http://

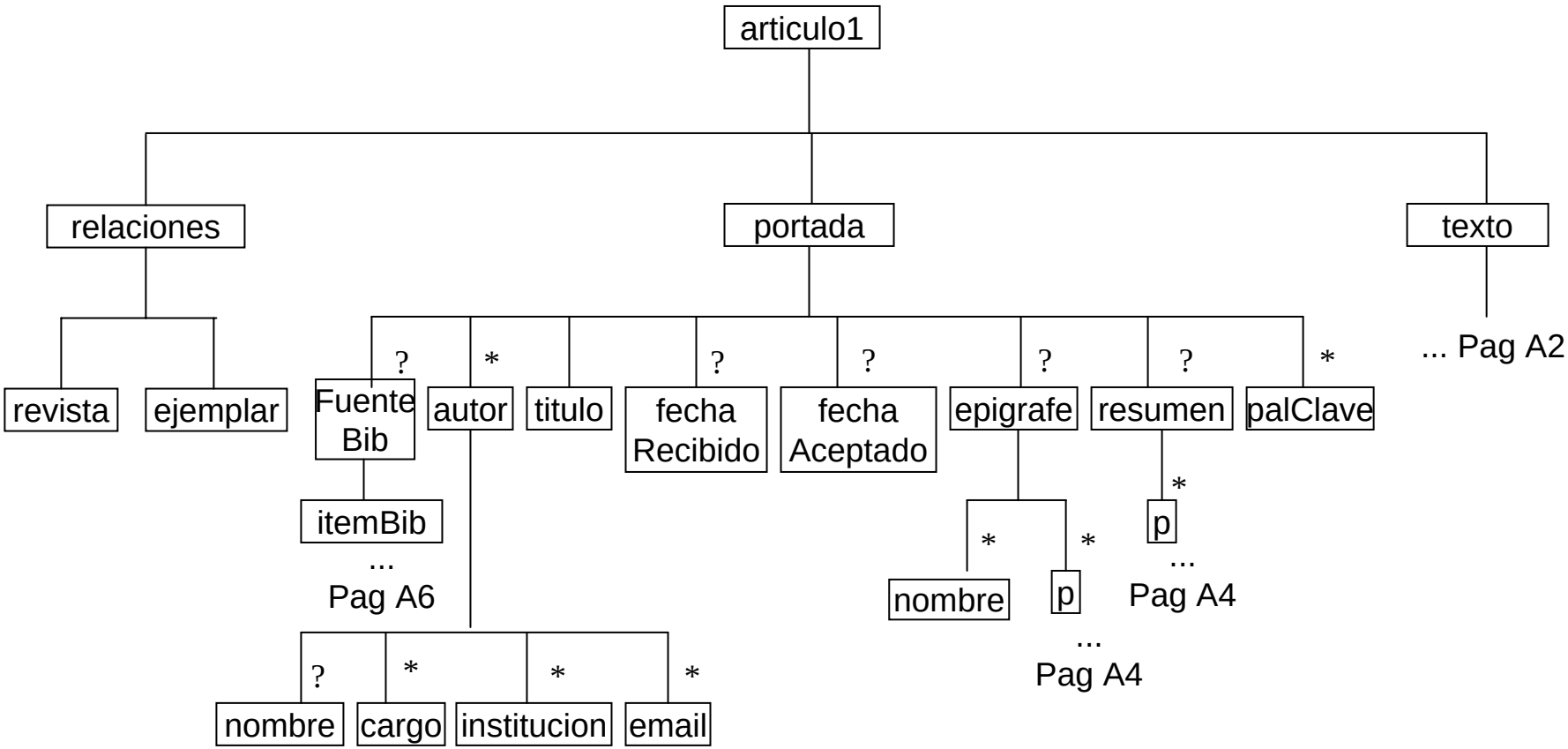
Troll, D. (2001). *How and Why Are Libraries Changing*? Recuperado de: http://www.diglib.org/use/whitepaper.htm

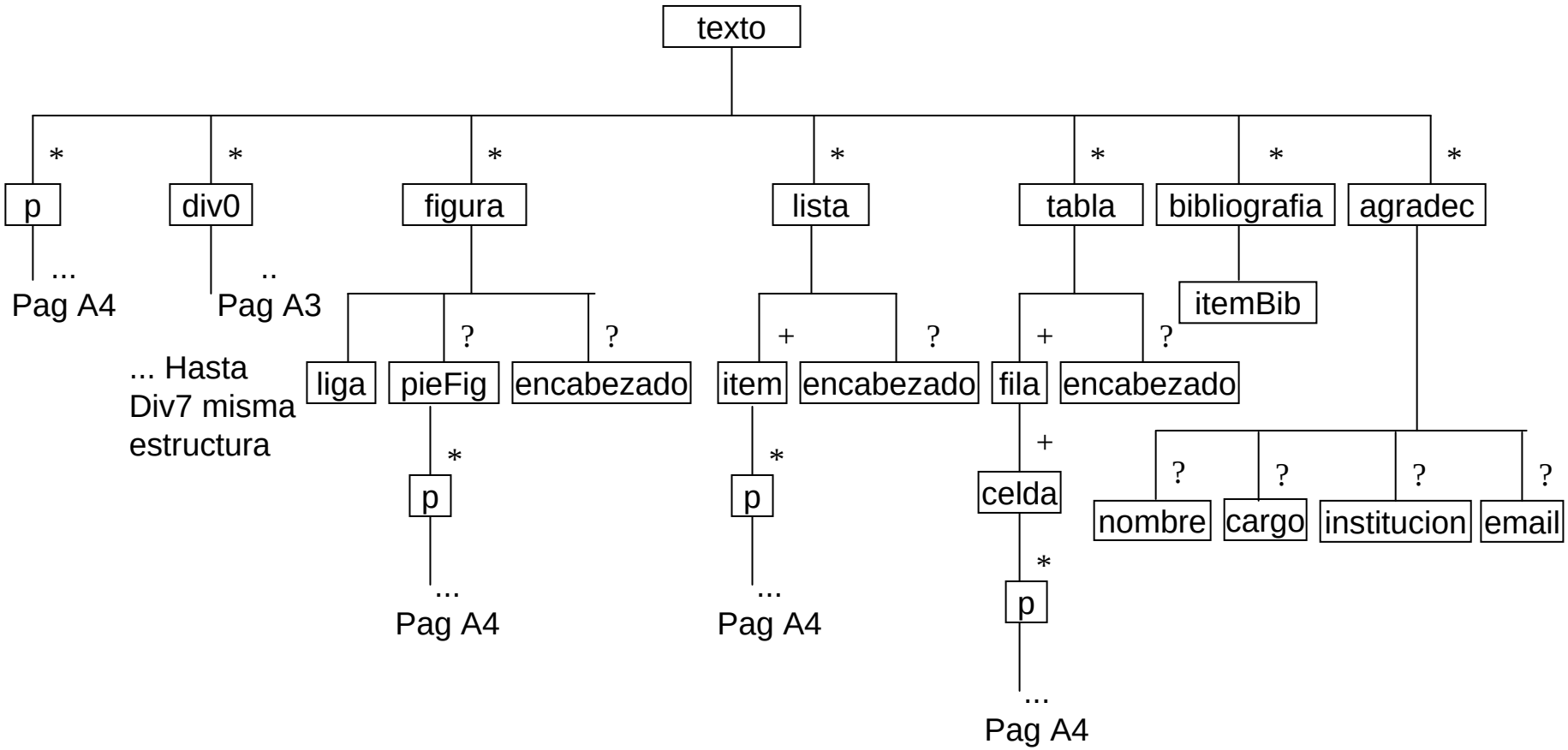
Universidad de Guadalajara (2001). Biblioteca : derechos de autor y propiedad intelectual. *Memorias VII Congreso de Bibliotecarios, Guadalajara, México*.

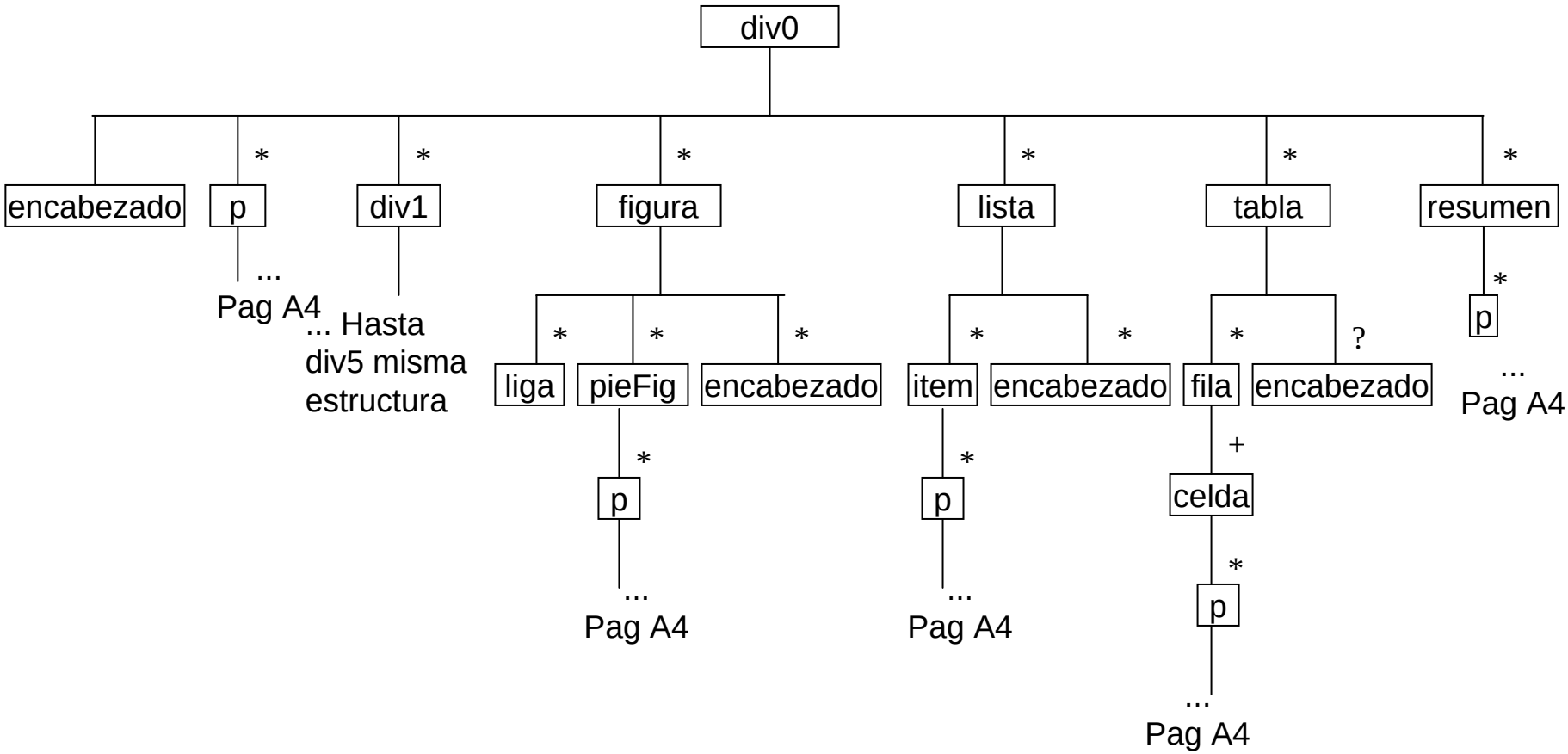
Wium, H. y Saarela J. (1999, octubre). Multipurpose web publishing using HTML, XML and CSS. *Communication of the ACM*. 95-101.

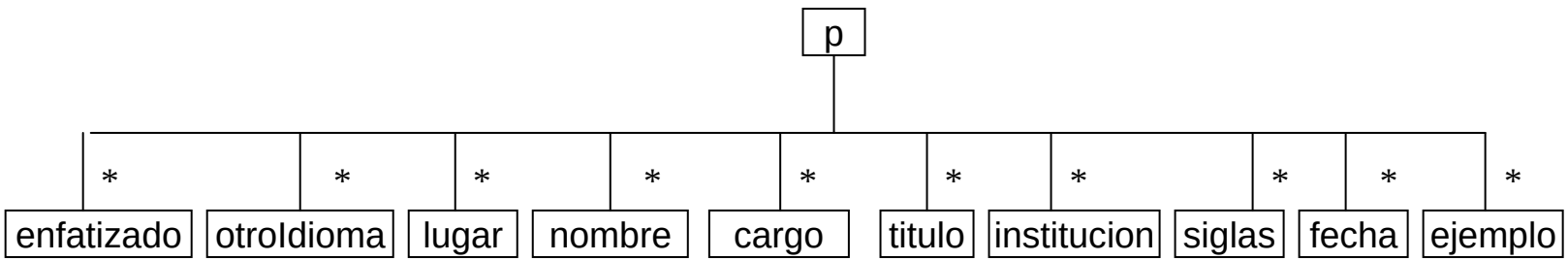
Wu, Y.D. y Liu, M. (2001). Content management and the future of academic libraries. *The Electronic Library*: 19(6): 432-439.

Anexo 1: Diagrama de Árbol de los Elementos del Vocabulario articulo1

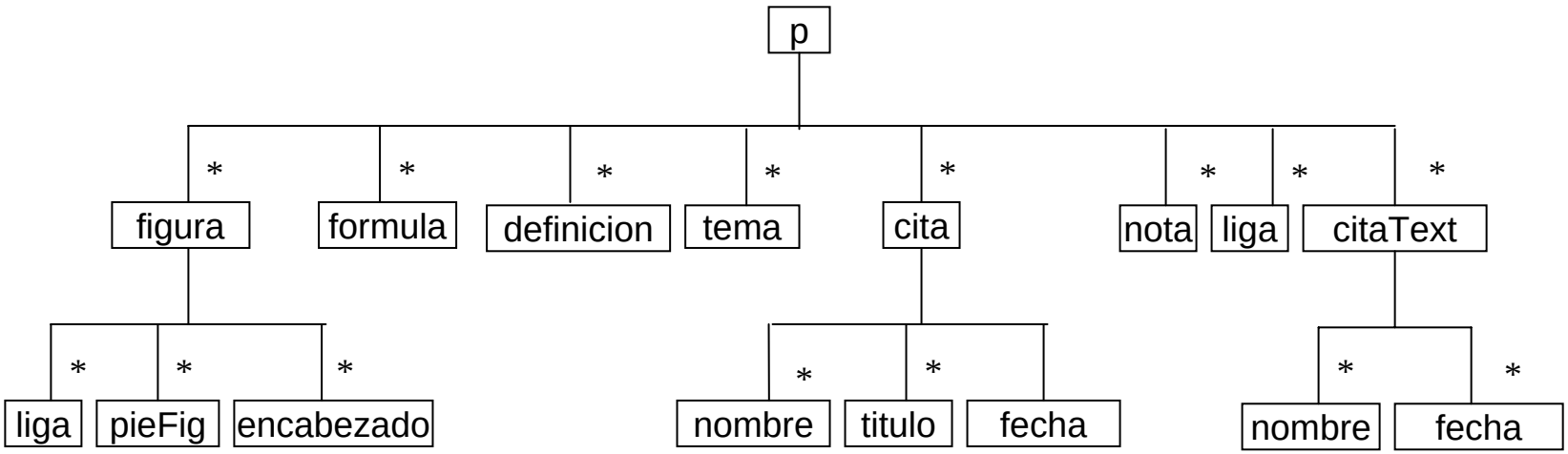



texto
p
div0
figura
lista
tabla
bibliografia
agradec
...
Pag A4
..
Pag A3
... Hasta
Div7 misma
estructura
liga
pieFig
encabezado
item
encabezado
fila
encabezado
itemBib
p
p
celda
nombre
cargo
institucion
email
...
Pag A4
...
Pag A4
p
...
Pag A4


div0
encabezado
p
*
...
Pag A4
div1
*
... Hasta
div5 misma
estructura
figura
*
liga
pieFig
encabezado
p
lista
item
encabezado
p
tabla
fila
encabezado
celda
p
+
?
resumen
p
...
Pag A4


p
*
enfatizado
*
otroIdioma
*
lugar
*
nombre
*
cargo
*
titulo
*
institucion
*
siglas
*
fecha
*
ejemplo

...
Pag A5


p
figura
formula
definicion
tema
cita
nota
liga
citaText
liga
pieFig
encabezado
nombre
titulo
fecha
nombre
fecha

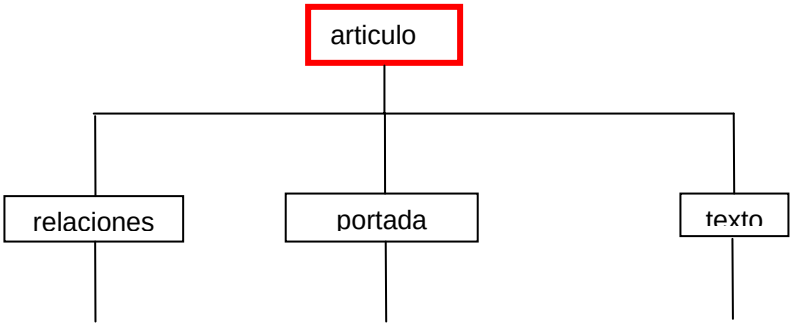
Anexo 2: Muestra de Hojas de Registro de Elementos del Vocabulario artículo

Forma de Elementos

Nombre del elemento: articulo_____ Número:01_____

Identificador Genérico: AR_____
Clase(s): _____

Modelo:



Contenido: Es el elemento raíz o root element y contiene todo la información del documento
_____
_____
_____
_____
_____

Justificación: Debe existir un elemento raiz y artículo es un buen nombre pues cada archivo contiene un artículo
_____
_____
_____
_____
_____
_____
_____
_____
_____

Es contenido en: _____
_____

Componentes relacionados: relaciones, portada y texto _____
_____

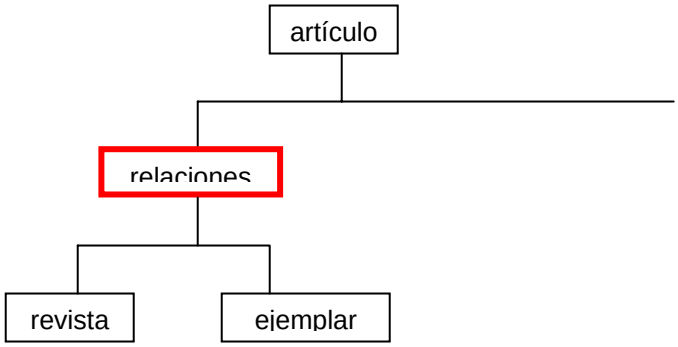
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea_____
_____
_____
_____

Forma de Elementos

Nombre del elemento: relaciones_____ Número: 02_____

Identificador Genérico: RE_____
Clase(s): _____

Modelo:



Contenido: Es el elemento que contiene las entidades revista y ejemplar, se asocian a través de un
ID._____
_____
_____
_____
_____
_____
_____
_____
_____

Justificación: Dado que los archivos de artículo no contienen información genérica en relación a la revista o especifica respecto del ejemplar o fascículo, se hace una referencia a esa información a traves de las ligas a las entidades correspondientes. _____
_____
_____
_____
_____
_____
_____
_____

Es contenido en: artículo - 01_____
_____

Componentes relacionados: revista y ejemplar _____
_____

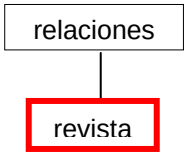
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea_____
_____
_____
_____

Forma de Elementos

Nombre del elemento: revista_____ Número: 03_____

Identificador Genérico: RE_____
Clase(s): _____

Modelo:



Contenido: El elemento revista contiene el título de la revista. En un modelo mas desarrollado podría incluirse un ID que relacionara con una entidad que posea toda la descripción bibliográfica del título de la revista _____
_____
_____
_____
_____
_____
_____
_____

Justificación: Todo artículo esta asociado a un título de revista
_____
_____
_____
_____
_____
_____
_____
_____
_____

Es contenido en: el elemento relaciones _____
_____

Componentes relacionados: relaciones _____ _____
_____

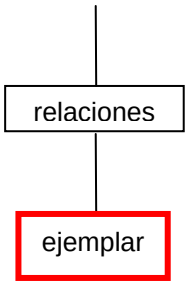
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea documentado el 13 de agosto de 2003_____
_____
_____
_____

Forma de Elementos

Nombre del elemento: ejemplar_____ Número:04_____

Identificador Genérico: EJ_____
Clase(s): _____

Modelo:



Contenido: El elemento ejemplar contiene la referencia al volumen y/numero de la revista que contiene el artículo. Se utilizan las abreviaturas n y vol. Por ejemplo para el numero 41 de una revista tenderemos en ejemplar <ejemplar>n 41 </ejemplar>

Justificación: Todo artículo esta asociado a un ejemplar

Es contenido en: relaciones _____

Componentes relacionados: relaciones, _____

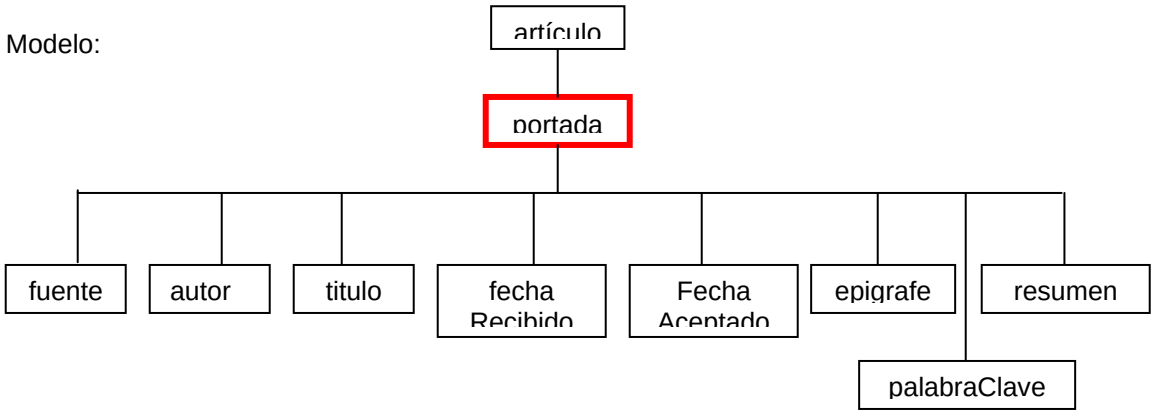
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea 13 de agosto de 2003 documentado__

Forma de Elementos

Nombre del elemento: portada_____ Número:05_____

Identificador Genérico: PO_____
Clase(s): _____

Modelo:



Contenido: Este elemento asocia todos los elementos que conforman la descripción bibliográfica del artículo, aquí se anidan los metadatos bibliográficos. A excepción de autor y titulo el resto de los elementos hijos son opcionales, solo se capturaran si existen en la fuente.

_____
_____
_____
_____
_____
_____
_____

Justificación: Los metadatos son elementos básicos para la recuperación de información del artículo y además lo identifican como único

_____
_____
_____
_____
_____
_____
_____

Es contenido en: el elemento raiz artículo_____
_____

Componentes relacionados: fuente, autor, titulo, fecha recibido,fecha aceptado, epígrafe, palabraClave y resumen_____
_____

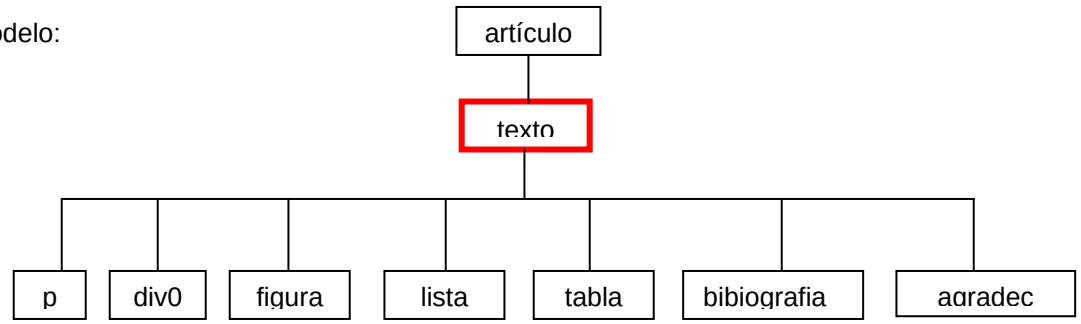
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 13 de agosto de 2003 documentación

Forma de Elementos

Nombre del elemento: texto_____ Número: 06_____

Identificador Genérico: TE_____
Clase(s): _____

Modelo:



Contenido: texto es el elemento que contiene la información sustancial del artículo, todas sus partes con opcionales .____
_____
_____
_____
_____
_____
_____
_____
_____

Justificación: Este elemento es fundamental pues contiene el artículo en si
_____
_____
_____
_____
_____
_____
_____
_____
_____

Es contenido en: el elemento raíz artículo
_____

Componentes relacionados: p, div0, figura, lista, tabla, bibliografía y agradec
_____

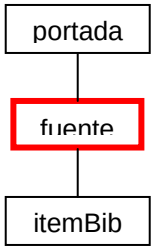
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 13 de Agosto de 2003 documentación _____
_____
_____
_____

Forma de Elementos

Nombre del elemento: fuente _____ Número: 07_____

Identificador Genérico: FU_____
Clase(s): _____

Modelo:



Contenido: Cuando un artículo está tomado de otra publicación este elemento registrara a través de sus elementos hijos todos los datos del item bibliográfico donde originalmente se publicó el material

Justificación: Aunque no es muy común existen casos de reproducciones de materiales originalmente publicados en otras fuentes, en especial el caso de las traducciones.

Es contenido en: portada _____

Componentes relacionados: itemBib _____

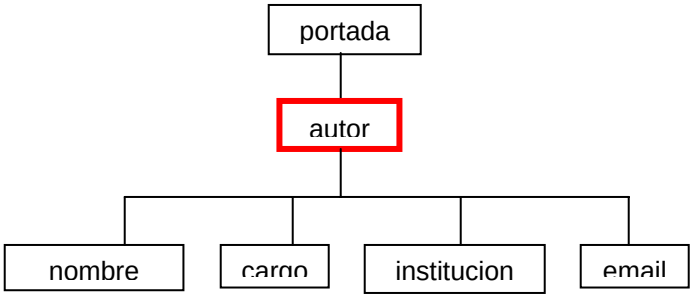
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, documentado 14 de agosto de 2003

Forma de Elementos

Nombre del elemento: autor_____ Número: 08_____

Identificador Genérico: AU_____
Clase(s): _____

Modelo:



Contenido: Corresponde a los autores del artículo, puede ser repetible y van asociado a cargo, institución y correo electrónico. En caso de que varios autores pertenecen a la misma institución solo se pondrá una vez.

Justificación: Debe existir un elemento raíz y artículo es un buen nombre pues cada archivo contiene un artículo

Es contenido en: portada _____

Componentes relacionados: nombre, institución, cargo e email___

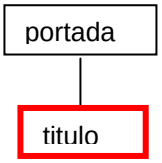
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentacion_____

Forma de Elementos

Nombre del elemento: titulo_____ Número:09_____

Identificador Genérico: _____
Clase(s): _____

Modelo:



Contenido: En este elemento se registra el titulo, subtitulo, títulos paralelos y cualquier elemento que pueda considerarse como parte del titulo.

Justificación: Todo artículo tiene un titulo y este es uno de los elementos mas importantes para la recuperación de contenido, sobre todo en los artículos de revistas científicas

Es contenido en: _____

Componentes relacionados: relaciones, portada y texto _____

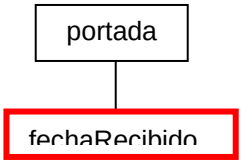
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea_____

Forma de Elementos

Nombre del elemento: fechaRecibido_____ Número:10_____

Identificador Genérico: FR_____
Clase(s): _____

Modelo:



Contenido: fechaRecibido registra la fecha en que el artículo fue recibido para su consideración de publicación

Justificación: Los lectores se pueden conocer la fecha en que el artículo fue recibido para publicación pues en revistas científicas la fecha en que los resultados de investigación fueron obtenidos es importante y no siempre coincide con la publicación.

Es contenido en: portada_____

Componentes relacionados: _____

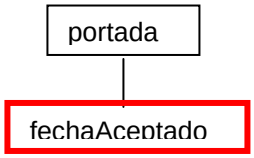
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, documentado 14 de Agosto de 2003

Forma de Elementos

Nombre del elemento: fechaAceptado_____ Número:11_____

Identificador Genérico: FA _____
Clase(s): _____

Modelo:



Contenido: fechaAceptado nos indica cuando fue aceptado el artículo por el consejo editorial de la revista. No todas las publicaciones indican este tipo de fechas por lo que este elemento es opcional

Justificación: Es un aspecto de la portada de tipo descriptivo, en lo referente a contenido podría indicar que el autor ha necesitado revisar poco su material si esta fecha y la fechaRecibido son cercanas o que el consejo editorial es muy flexible en sus criterios de revisión.

Es contenido en: portada _____

Componentes relacionados: _____

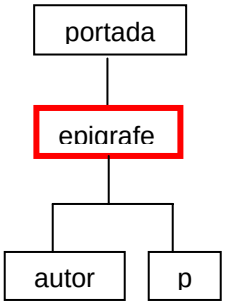
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, documentado 14 de Agosto de 2003 _____

Forma de Elementos

Nombre del elemento: epigrafe_____ Número:12_____

Identificador Genérico: EP _____
Clase(s): _____

Modelo:



Contenido: Frase sentencia o cita que se coloca al inicio de un escrito sugiriendo algo de su contenido, o lo que lo ha inspirado.

Justificación: Es parte del texto y en muchas ocasiones da indicaciones del contenido del texto

Es contenido en: portada_____

Componentes relacionados: autor y párrafo _____

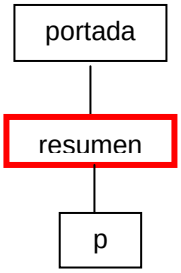
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentación

Forma de Elementos

Nombre del elemento: resumen_____ Número:13_____

Identificador Genérico: RE_____
Clase(s): _____

Modelo:



Contenido: Este elemento contiene la exposición resumida del contenido del artículo, algunas revistas los traen.
_____
_____
_____
_____
_____
_____
_____
_____
Justificación: El resumen contiene usualmente los principales elementos de contenido
_____
_____
_____
_____
_____
_____
_____
_____
_____
Es contenido en: portada_____
_____

Componentes relacionados: párrafo_____
_____

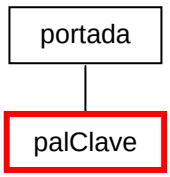
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentación _____
_____
_____

Forma de Elementos

Nombre del elemento: palClave_____ Número:14_____

Identificador Genérico: PC_____
Clase(s): _____

Modelo:



Contenido: Palabras claves, también conocidas como keywords registradas en el artículo.
_____
_____
_____
_____
_____
_____
_____
_____
_____
Justificación: Son fundamentales para la recuperación de contenidos
_____
_____
_____
_____
_____
_____
_____
_____
_____
Es contenido en: portada _____
_____

Componentes relacionados: _____
_____

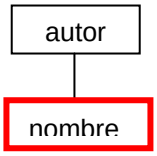
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentación_____
_____
_____
_____

Forma de Elementos

Nombre del elemento: nombre_____ Número:15_____

Identificador Genérico: NO_____
Clase(s): _____

Modelo:



Contenido: El nombre incluye todos lo referente a primer nombre, segundo nombre, apellidos y cualquier inicial del autor. No incluye el tratamiento como Doctor, Maestro, etc.

Justificación: Para efectos de la indización por palabras no hace falta separar nombre de apellido.

Es contenido en: autor_____

Componentes relacionados: _____

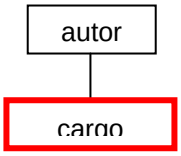
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de Agosto de 2003 documentación _____

Forma de Elementos

Nombre del elemento: cargo_____ Número:16_____

Identificador Genérico: CA_____
Clase(s): _____

Modelo:



Contenido: La función o empleo del autor, si es que se consigna en la publicación original
_____
_____
_____
_____
_____
_____
_____
_____
_____
Justificación: Es común registrar la responsabilidad laboral de un autor en los artículos
_____
_____
_____
_____
_____
_____
_____
_____
_____
Es contenido en: autor_____
_____

Componentes relacionados: _____
_____

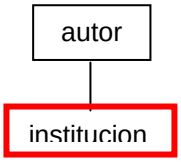
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentación _____
_____
_____
_____

Forma de Elementos

Nombre del elemento: institución _____ Número:17_____

Identificador Genérico: IN_____
Clase(s): _____

Modelo:



Contenido: insitución a la esta adscrito el autor (es) del artículo y que se encuentra registrado en la publicación.

Justificación: Se registra como parte de los metadatos en relación al autor

Es contenido en: autor_____

Componentes relacionados: _____

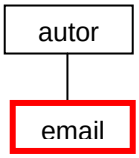
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentación

Forma de Elementos

Nombre del elemento: email_____ Número:18_____

Identificador Genérico: EM_____
Clase(s): _____

Modelo:



Contenido: Registra el correo electrónico del autor, a través del cual los lectores pueden comunicarse con el mismo.

Justificación: Es parte de los datos asociados al autor y suele incluirse en las publicaciones de artículos científicos

Es contenido en: autor_____

Componentes relacionados: _____

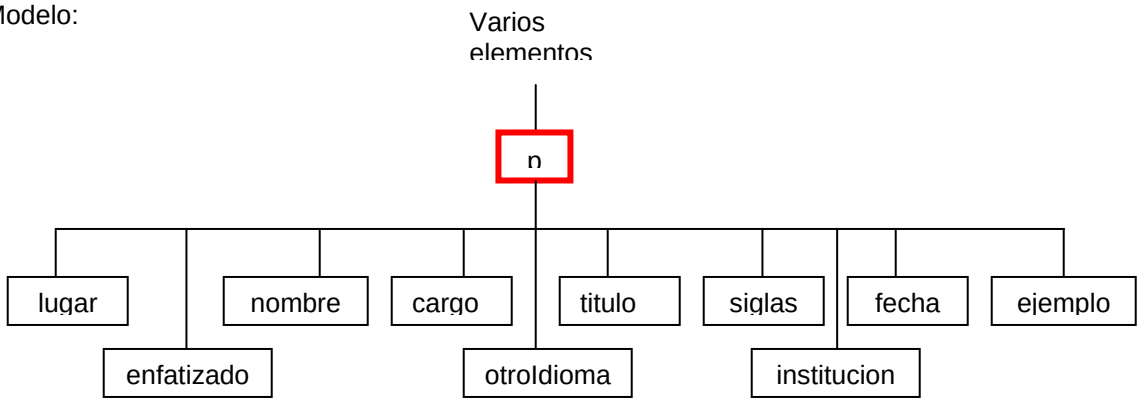
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentacion _____

Forma de Elementos

Nombre del elemento: p_____Número:19_____

Identificador Genérico: PA_____
Clase(s): _____

Modelo:



Contenido: El párrafo registra los conjuntos de palabras separados por un punto y aparte.

Justificación: El párrafo es un elemento fundamental en la redacción narrativa utilizada en los artículos. Un párrafo suelo contener una idea completa.

Es contenido en: Varios elementos como: texto, epígrafe, resumen

Componentes relacionados: enfatizado, otroIdioma, lugar, nombre, cargo, titulo, insititucion, siglas, fecha, ejemplo

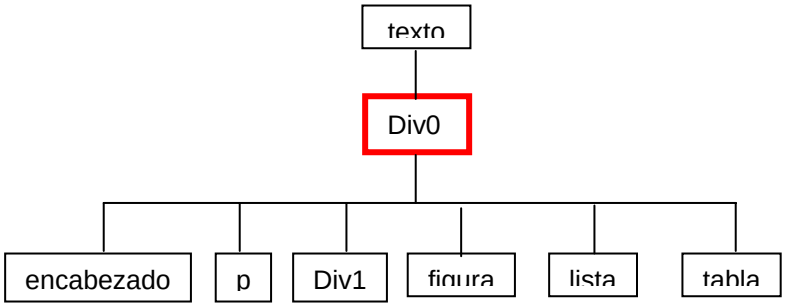
Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de agosto de 2003 documentacion _____

Forma de Elementos

Nombre del elemento: div0_____ Número:20_____

Identificador Genérico: D0_____
Clase(s): _____

Modelo:



Contenido: Este elemento permite divisiones de partes de un artículo, el elemento div0 es el primer nivel pero se admiten hasta 5 niveles, aunque lo mas común es utilizar máximo haste 3 niveles

Justificación: Los artículos están estructurados en partes tales como introducción o conclusiones y el marcado de divisiones sirve para hacer búsquedas mas especificas de contenido.

Es contenido en: texto _____

Componentes relacionados: encabezado, div, p, figura, lista, tabla, resumen_____

Historia de creación y cambio: mayo 2003 idea, 28 de julio de 2003 idea, 14 de Agosto de 2003 documentacion _____