2. ANTECEDENTES Y MARCO TEÓRICO

Esta investigación se enmarca en la convergencia de varios temas: bibliotecas digitales, digitalización, 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 que ataca 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 alta 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.

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:

---

[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 capítulo 6: Bibliografía.

[2] Al respecto 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 (Confererencia conjunta sobre el tema de bibliotecas digitales organizada por la ACM y la IEEE).

- 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.

Se puede ver hoy día que los enfoques cada vez se juntan más y se observan desarrollos y propuestas con visión multidisciplinaria; lo anterior genera aportaciones más ricas 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 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

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

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

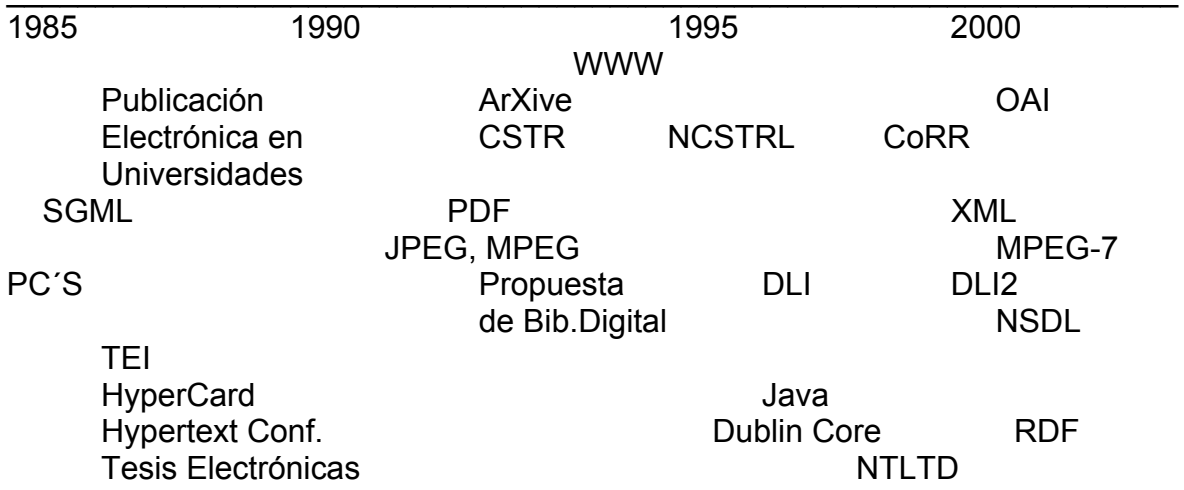

Figura 2.1. Cronología de tecnologías relacionadas con bibliotecas digitales, tomado de Fox, E. y Urs, S.R., 2002, p.508.

La figura anterior 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 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. Una vez 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 y de la IEEE[5], que a partir de 2000 se llevan a cabo como una sola conferencia, y la reunión de la American Society for Information Science and Technology.[6], pero debe considerarse que estos desarrollos se han dado, aunque en diversa medida, en el mundo entero.

Las novedades sobre biblioteca digital se han publicado inicialmente en revistas[7] 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

[5] http://www.jcdl2004.org
[6] http://www.asis.org
[7] D-Lib Magazine http://www.dlib.org

Libraries[8] 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 precisa a través del uso de marcado XML está incluido en el punto 7.

Como resultado de la globalización tenemos que en las conferencias hay presencia mundial, y personalidades como Edward Fox o Ian Witten[9] pueden encontrarse en reuniones efectuadas en diferentes continentes.

Nueva Zelanda ha sido una región que ha generado propuestas valiosas 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 Naciontes Unidas para el fomento de la Educación, la Ciencia y la Cultura) en los países en

[8] http://www.ecdl2003.org
[9] 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.

desarrollo y particularmente en Latinoamérica[10]. 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.

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

[10] El Mtro. Claudio Menezes de la Oficina Regional de Ciencia para América Latina, UNESCO proporcionó el software Greenstone en CD durante la conferencia II Conferencia Internacional

- Cooperación internacional.

Una experiencia de biblioteca digital cercana a Latinoamérica es el sitio español llamado Cervantes Virtual[11], 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[12].

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[13].

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)[14]
- Tecnológico de Monterrey (ITESM-Monterrey)[15]
- Universidad de Colima[16]
- Universidad Autónoma de México (UNAM-DGSCA)[17]

---

sobre Bibliotecas Universitarias en la UNAM, octubre 2003.

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

[12] La Universidad Iberoamericana tiene un convenio de colaboración con este proyecto desde 2001. Si bien se han enviado contenidos a ser publicacos 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.

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

[14] http://biblio.pue.udlap.mx/digital/desarrollo.html

[15] http://copernico.mty.itesm.mx/~tempo/Proyectos/

- Instituto Politécnico Nacional (IPN)[18].

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 Cómputo (CIC), 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

[16] http://bdigital.ucol.mx/Menu.htm

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 metodogías 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
5. Producción digital
    a. Escaneo
    b. OCR[19]
    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

---

[17] http://www.bibliodgsca.unam.mx/
[18] http://www.cic.ipn.mx/laboratorios/coord3.htm

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[20], 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 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

---

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

proyectos y un elemento totalmente humano: el liderazgo[21], sin el cual se han visto no pocos fracasos.

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[22], la única manera de publicarlo en formato digital es obteniendo los derechos explícitamente por parte de los titulares, en caso contrario la 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 material antiguo[23], lo cual no siempre es la solución adecuada 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

---

[20] 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.

[21] 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.

[22] 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

[23] 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).

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)[24], 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 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.

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

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[25] [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"

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.

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

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 padre 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) manifiestaba 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 3 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 uno los metadatos son genéricamente datos sobre los datos. El concepto varía de una disciplina a otra (Rivera Aguilera, 2001)[26], en el contexto de bibliotecas digitales se refiere a los datos que describen un objeto 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.[27]

---

[26] 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.

[27] 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

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 covirtió 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)[28] tienen como principal referencia al SGML, esta tecnología define más que

[28] 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.

una forma de archivar información un metalenguaje para generar aplicaciones o vocabularios de marcado específicos con una preocupación fundamentalmente estructural[29], 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
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.

Más lamentable aún es que no haya más publicaciones sobre el tema en español que enriquezcan la discusión.

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.[30]

---

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

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

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 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 Educacion Enrique Jose 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 Comunicacion y Educacion
          como dos procesos complejos y que estan muy estrechamente relacionados. Se enuncian
          algunos aspectos esenciales que conforman una definicion integral de Comunicacion. Se
          valora el impacto de ambas categorias para el proceso pedagogico, donde se concibe la
          Comunicacion 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 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>
    </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>

…….  Definción 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[31]

"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. Afortundamente, 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)[32] 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

[31] 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.

[32] URI engloba a los URL y URN. Los primeros son los más comunes direcciones de recursos en internet y son dependientes de la ublicación, 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 bibliotecaria 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 puede desplegarse a través de las hojas y lenguajes 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[33]. 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

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

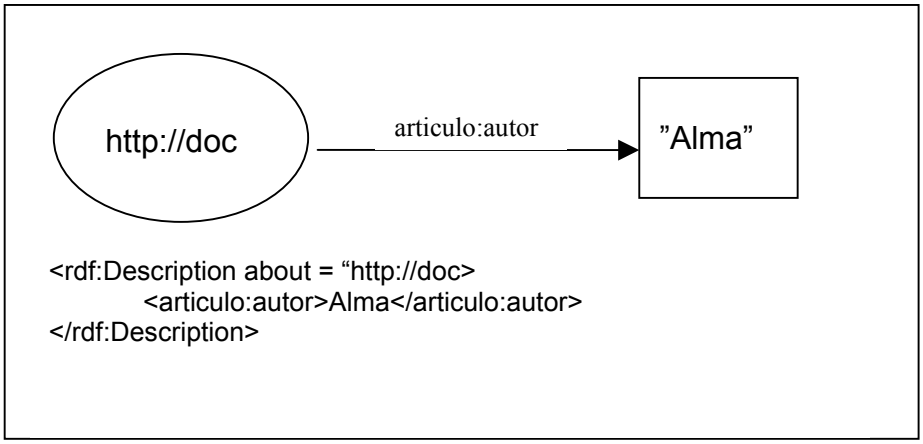

Figura 2.12 Modelo RDF. (tomado en parte de Miller, E. y Hillman, D., 2002, p.60)

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 incio del archivo. Si el esquema referido a tarjetas de presentación virtuales

(xmlns:vcard="http://www.imc.org/pdi/Vcard/") hubiése sido integrado en ese ejemplo el archivo pudiése 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 refiera 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**[34] 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)[35]. 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.

[34] http://uk.dublincore.org/schemas/xmls/, utilzado como metadatos en muchos bibliotecas digitales
[35] 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)[36] 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

[36] 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)[37] /**METS** (Metadata Encoding and Transmission Standard)[38]. 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

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

- **SciELO** (Scientific Electronic Library Online)[39]. Este esquema es una propuesta brasileña generado en BIREME (Centro Latinoamericana y del 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[40]

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

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

despliegue más que para recuperación[41] y que de alguna manera poseen mayor impacto pues este tipo de empresas son las que están generando mayores repositorios de artículos y llegando a más cantidad de usuarios a través de sus paquetes de colecciones de revista como Academic Search Premier[42] o Academic Literature[43]

Una vez revisados los vocabularios tanto descriptivos como los que combinan descripción y estructura, debe señarlarse 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).

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

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
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 subdisciplinas 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)[44]:

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 posiblidades 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.
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

[44] El tema de los índices y buscadores que los aprovechan es enorme y no se tratará en detalle en esta investigación ya que se aprovecharán herramientas disponibles de software gratuito para las

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.

---

pruebas con el marcado estructural propuesto.

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[45]. 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 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

[45] 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.

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[46] 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

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

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.

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 bloqueda 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 especìfico de este trabajo pueden consultarse en la extensa bibliografía sobre el tema que instituciones como el INASP (Internacional Network for the Availability of Scientific Publications) tienen disponible.[47]

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

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

marcado, catalogación, bases de datos y finalmente, pero no menos importante la publicación electrónica de revistas.