

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

E.A.P. DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

MUSUJ QUIPU: Metodología para la preservación y publicación de documentos digitales, aplicado a la Biblioteca Central Pedro Zulen de la UNMSM.

Tesis para optar el Título Profesional de Ingeniero de Sistemas.

AUTORA:

Yanet Martínez Berrocal

ASESORA: Dra. Nora la Serna Palomino

LIMA – PERÚ 2007

Dedicatoria

A la naturaleza por ser fuente de vida.

A los niños por ser la continuidad de nuestra raza, cultura, costumbre y creencia.

A mi pátria querida, Perú por ser cuna de hombres justos e ingeniosos.

A mis padres Emiliano Martínez y Otilia Berrocal,

a mis hermanos Bladimir, Madler y Alan, a mi abuelita Benedicta,

a mi esposo Arturo Luna, por ser el centro de mi vida, motivarme, comprenderme y servirme de ejemplo con cada uno de sus actos.

Agradecimientos

Agradezco a mi alma Mater la Universidad Nacional Mayor de San Marcos por abrirme sus puertas y permitir ampliar mis conocimientos y por darme la oportunidad de conocer a muchas personas importantes en mi vida.

Agradezco a los profesores de la facultad de Ingenieria de Sistemas por brindarme sus conocimientos, servirme de ejemplo e incentivar me hacia la investigación.

Agradezco a los administrativos, alumnos y a todos los integrantes de la facultad de Ingenieria de Sistemas por hacerme sentir en estos cinco años, como en casa.

Agradezco a la unidad de informática de la Biblioteca Central Pedro Zulen por acogerme y permitir realizar mis prácticas durante 3 años.

Agradezco a los que crearon y difundieron el proyecto Cybertesis, que me sirvió de fuente de orientación en los momentos de incertidumbre.

Agradezco por su interés y colaboración en la realización del sistema Musuj Quipu, a las siguientes personas: Patricia Salazar, Liceth Mancisidor, Carlos Laban, Denisse Muñante y Paul Lorenzo. Y mis reconocimientos especiales al Sr. Libio Huaroto y a mi asesora Dra. Nora la Serna Palomino que han leído y han colaborado con sus sugerencias y sus conocimientos.

Agradezco a todos mis amigos por su colaboración, apoyo y entusiasmo, que me incentivaron para la culminación de este trabajo.

Contenido

Capítulo I:

Introducción	9
---------------------------	----------

Capítulo II:

Planteamiento del problema	14
---	-----------

2.1. Definición del Problema.....	14
-----------------------------------	----

2.2.Objetivos de la tesis	15
---------------------------------	----

2.2.1. Objetivo general	15
-------------------------------	----

2.2.2. Objetivos específicos.....	15
-----------------------------------	----

2.3. Justificación del Proyecto.....	16
--------------------------------------	----

2.4. Alcance	17
--------------------	----

Capítulo III:

Bibliotecas Digitales	18
------------------------------------	-----------

3.1. Definición.....	18
----------------------	----

3.2. Antecedentes y evolución del concepto de bibliotecas digitales	21
---	----

3.3. Características.....	25
---------------------------	----

3.3.1. Contenidos	28
-------------------------	----

3.3.2. Preservación.....	29
--------------------------	----

3.3.3. Flexibilidad	30
---------------------------	----

3.3.4. Cooperación.....	30
-------------------------	----

3. 3.5. Tecnología	30
--------------------------	----

3.3.6. Legislación	31
--------------------------	----

3.3.7. Formatos.....	31
3.4. Metadato.....	32
3.4.1. Antecedentes y definición.....	33
3.4.2. Convenciones, sistemas de metadatos.....	36
3.5. Problemática de bibliotecas digitales	38
3.5.1. Cambios tecnológicos	39
3.5.2. Cambios económicos	40
3.5.3. Cambios sociales y culturales.....	40
Capítulo IV:	
Preservación digital.....	44
4.1. Introducción.....	44
4.2. Objetivo	47
4.3. Herencia y Patrimonio Digital	46
4.4. Desafíos	49
4.5. Estrategias técnicas	51
4.6. Estrategias organizacionales	56
Capítulo V:	
Revisión de estándares.....	58
5.1. Introducción.....	58
5.2. Estándares para transferencia de datos	58
5.2.1. Lenguajes de marcado.....	58
5.2.2. SGML (Standard Generalized Markup Language, ISO 8879)	59
5.2.3. Lenguaje de Marcado para Texto	60
5.2.4. Lenguaje de Marca Extensible.....	60
5.2.4.1 Estructura de un documento XML.....	62

5.2.4.2. Diferencias entre XML Y HTML	64
5.2.4.3. Declaración de Tipo de Documento (DTD)y Esquema	66
5.3. Estándares de metadatos	67
5.3.1. Estándares World Wide Web Consortium(W3C)	67
5.3.2. Estándar de metadatos Dublin Core (DC)	68
5.4. Iniciativa de texto codificado (text encoding initiative TEI)	74
Capítulo VI:	
Recuperación de Información.....	80
6.1. Introducción.....	80
6.2. Almacenamiento y recuperación de información en datos no estructurados	81
6.2.1. Indexación de un documento.....	81
6.2.2. Modelo de recuperación de información.....	85
6.3. Lucene.....	89
6.3.1. Indexando con Lucene	90
6.3.2. Resumen de clases que están implementadas en el Api de Lucene.....	91
6.3.3. Servidor “Seekafile”	92
Capítulo VII:	
Metodología Musuj Quipu	95
7.1. Introducción.....	95
7.2. Descripción de la Metodología Musuj Quipu.....	96
7.2.1. Marcación del Documento.....	97
7.2.2. Estructura de Carpetas	100
7.2.3. Uso de Metadato	103

7.2.4. Proceso de conversión de Archivos.....	103
7.2.5. Proceso de Indexación del contenido y del metado.....	104
7.2.6. Proceso de búsqueda y recuperación.....	104
7.3. Análisis y diseño del sistema Musuj Quipu.....	104
7.3.1. Actores del sistema	105
7.3.2. Diagrama de casos de uso general	106
7.3.3. Modelo de datos para el Contenido del Libro	107
7.3.4. Modelo de datos del Metadato	108
7.3.5. Arquitectura de Musuj Quipu UNMSM	108
7.4. Principales módulos del sistema	111
7.5. Implementación del sistema Musuj Quipu	113
7.5.1. Selección de tecnologías	113
7.5.2. Especificaciones Técnicas.....	114
7.5.3. Desarrollo del Sistema Musuj Quipu	115
7.5.3.1. Componentes del Sistema Generador.....	115
7.5.3.2. Formato del documento XML con el Estándar TEI.....	118
7.5.3.3. Componentes de la aplicación de búsqueda de datos en el cliente.....	121
7.6. Descripción del trabajo desarrollado	123
Capítulo VIII:	
Análisis de los resultados obtenidos y trabajos futuros	126
8.1. Análisis de los resultados	126
8.1.1. Consistencia de base datos	126
8.1.2. Análisis del tiempo de procesamiento del documento digital.....	127
8.1.3. Análisis de tiempo de búsqueda.....	129

8.2. Trabajos futuros.....	131
Capítulo IX:	
Conclusiones	132
Glosario	134
Referencias bibliográficas	140
Anexos	149
Anexo 1: Acuerdo de Cartagena	149
Anexo 2: Resumen de modelos de recuperación de información [56]	153
Anexo 3: Hojas de estilo XSL creados para la implementación del sistema.	164
Anexo 4: Guía de usuario del Sistema Generador, módulo de ingreso de datos.....	175
Anexo 5: Guía de usuario del Sistema Musuj Quipu	184

Resumen

Actualmente la información es uno de los recursos más valorados pero nos enfrentamos a grandes volúmenes de ellas, cuyo crecimiento es exponencial a raíz de la expansión de internet y el avance tecnológico. Para su almacenamiento, acceso y procesamiento se usan diversas tecnologías, quedando estas obsoletas principalmente por los siguientes motivos: políticas organizacionales, el cambio, el tiempo y el uso por grupos minoritarios; en consecuencia la información que usa dichas tecnologías se vuelve inaccesible, por ello nace la necesidad de preservar y resguardar los datos. Principalmente las bibliotecas digitales se ven afectados con el problema debido a decisiones centradas en la selección de tecnologías y tipos de conversión usados para publicar los libros digitales.

En este trabajo de tesis se plantea una metodología para la preservación, publicación y recuperación de los libros en formato digital, usando los estándares internacionales como la iniciativa de texto codificado (TEI), el Dublin Core (DC), y tecnologías de archivos de formatos abiertos como el XML y HTML. Se inicia describiendo la sustentación teórica y los estándares de publicación de documentos digitales existentes, los cuales van a sustentar el desarrollo del prototipo de la metodología planteada. Esta se compone de cinco fases: Concepción del problema y análisis de viabilidad, Investigación de los estándares, Normas y tecnologías, Proceso de implementación y desarrollo del prototipo e Implantación del prototipo.

Palabras claves: Biblioteca digital, Digitalización, Preservación digital, XML, Dublin Core, TEI, Estructura de datos, Lucene.

Abstract

Currently, the information is one of the most valuable resources; we have lots of volumes of it that grow exponentially as a result of the Internet expansion and the technological advance. For their storage, access and processing we used several technologies, but these become obsolete by the following reasons: enterprise policies, the change, the time spending and the use by minority groups; consequently the information that uses these technologies returns inaccessible, that's why the need to preserve and to protect the data is born. Mainly the digital libraries are affected with the problem due to decisions centered in the selection of technologies and types of conversion used to publish digital books.

In this thesis work one considers a methodology for the preservation, publication and recovery of books in digital format, using the international standards like the codified text initiative (TEI), Dublin Core (DC), and technologies of archives of opened formats like XML and HTML. One begins describing the theoretical sustentation and the existing digital document publication standards, which are going to sustain the development of the prototype of the raised methodology. This is made up of five phases: Conception of the problem and the analysis of viability, Investigation of the standards, Norms and technologies, Implementation Process and Prototype's Development and the Prototype's Implantation.

Key words: Digital library, Digitalization, Digital Preservation, XML, Dublin Core, TEI, Structure of data, Lucene.

Capítulo I: Introducción

La UNESCO ha reconocido la importancia del problema de conservación y preservación de contenidos digitales y ha escrito en uno de sus artículos lo siguiente : “El patrimonio digital del mundo corre el peligro de perderse para la posteridad. Contribuyen a ello, entre otros factores, la rápida obsolescencia de los equipos y programas informáticos que le dan vida, las incertidumbres existentes en torno a los recursos, la responsabilidad y los métodos para su mantenimiento y conservación y la falta de legislación que ampare estos procesos”[65].

El surgimiento de las bibliotecas digitales a finales del siglo pasado se debe a la gran demanda de información que exige el mundo moderno, convirtiendo la información en uno de los recursos más valorados, por ende es importante garantizar su preservación y el acceso a ella desde el punto de vista cualitativo como cuantitativo, esto implica ciertos cambios sociales, económicos, organizacionales y principalmente tecnológicos.

Por otro lado, hoy en día también ha cambiado la forma de transmitir la información y el conocimiento, desde épocas antiguas el conocimiento ha sido transmitido de generación en generación a través de los libros como objeto físico, que está compuesto por átomos, la cual se tenía que imprimir, empastar, conservar y distribuir. Actualmente sucede lo mismo en el mundo digital que se ha creado, el cual moviliza muchos recursos diferentes a los tradicionales. Al pasar los átomos a bits se va cambiando la forma de manejo

del contenido y el acceso a la misma, se requiere el uso de la tecnología y un cambio de paradigma por ello lo que sigue es poder implementar en la práctica educativa una biblioteca digital que cumpla con ese fin, de transmitir conocimiento y ser nueva fuente de información para las siguientes generaciones. Cabe recalcar que las bibliotecas digitales no van a desplazar totalmente a las bibliotecas reales en su forma si no, mas bien es un avance en cuanto al tratamiento y difusión de la información ya que es más fácil transmitirlo en esta forma, en la cual no se podrá deteriorar, perder, romper, borrar, etc. y principalmente llegará a mayor cantidad de persona en lugares inimaginables. Los libros y bibliotecas digitales abren nuevas esperanzas para las instituciones educativas más pobres y marginadas, modificará decisivamente las inversiones culturales del individuo y de la sociedad.

La actualización e inserción en la producción y difusión del conocimiento es de suma importancia por ello con este fin, este trabajo de tesis intenta aportar a la preservación y difusión del conocimiento a través de la Biblioteca Digital Pedro Zulen de la UNMSM, implementando una metodología de trabajo que consta de: técnica de preservación, arquitectura de datos y recuperación de información con el objetivo de preservar el patrimonio digital y facilitar el proceso de generación de documentos digitales para que nuestra casa de estudios pueda llegar a publicar mayor cantidad de material bibliográfico en Internet y a la vez conservarlo para las futuras generaciones.

La educación y la comunicación son los medios más seguros para transmitir información. La tecnología ha creado ambientes que le permiten interactuar, transportar y difundir el conocimiento a los lugares más recónditos del planeta, pero la tecnología es cambiante, por ello la preservación de los materiales digitales no esta asegurada por sí sola, si no

se tiende a usar normas y estándares para lograr preservarlos y de esta manera asegurar su perdurabilidad en el tiempo.

El contenido de este trabajo de tesis está desarrollado en nueve capítulos, se inicia con la *introducción*, continuando con el *planteamiento del problema y justificación*, donde se explica el problema existente en la biblioteca Pedro Zulen, se define los objetivos y se plantea la solución, en los capítulos III, IV, V y VI se describe el marco teórico que es base para el desarro de la solución propuesta, los temas detallados son:

Bibliotecas Digitales (donde se define el concepto y evolución de bibliotecas digitales), *Preservación Digital* (en este capítulo se define el concepto de preservación digital, historia y se describe proyectos existentes sobre preservación del patrimonio digital), *Revisión de estándares* (en este capítulo se revisan diversos estándares para el tratamiento de documentos digitales) y en *Recuperación de Información* (se describe las teorías de sistemas de recuperación de información existentes). En el capítulo VII se describe la *Metodología Musuj Quipu* y se explica detalladamente los procesos realizados, seguidamente en capítulo VIII se *analizan los resultados obtenidos y se propone los trabajos futuros* a realizarse y finalmente en el capítulo IX se propone las *conclusiones*.

Capítulo II: Planteamiento del problema

2.1. Definición del Problema

Actualmente la Biblioteca Central Pedro Zulen cuenta con una biblioteca digital, la cual contiene diferentes tipos de colecciones (libros, revistas, tesis, monografías, etc.). Se observa que no se cuenta con una metodología adecuada que facilite el procesamiento y almacenamiento de los recursos bibliográficos digitales trayendo como consecuencia la demora en la publicación, su difícil acceso y recuperación del material.

Los procesos actuales para digitalización, publicación y difusión de información no se ajustan a ninguna norma internacional para la preservación del patrimonio digital, ni cuentan con herramientas informáticas apropiadas para la aplicación de estos estándares. Es por ello que se hace necesaria la implementación de procesos que garanticen su accesibilidad en el tiempo.

Los procesos actuales de digitalización requieren un determinado número de personas y tiempo, lo cual trae como consecuencia un elevado presupuesto, y una menor producción de contenidos digitales.

2.2. Objetivos de la tesis

El objetivo de este trabajo de tesis se divide en: objetivo general y objetivos específicos, los cuales se detallan a continuación.

2.2.1. Objetivo general

Tenemos como objetivo general los siguientes:

1. Proponer y desarrollar una metodología de trabajo para la preservación, publicación y recuperación de documentos digitales aplicado a la Biblioteca Central Pedro Zulen de la UNMSM.
2. Mejorar el funcionamiento de la biblioteca digital que la universidad posee, brindando una herramienta de software creado en base a la metodología propuesta, que facilite la aplicación de estándares y que permita la optimización de los procesos de digitalización.

2.2.2. Objetivos específicos

- Investigar sobre las diversas iniciativas de preservación digital y su aplicación en la bibliotecas digitales y luego seleccionar las mas óptimas.
- Levantar información, analizar, diseñar , desarrollar e implantar el prototipo para preservación de información de acuerdo a los estándares internacionales seleccionados .
- Digitalizar los documentos de forma estructurada y estandarizada.
- Disminuir los tiempos en los procesos de digitalización y publicación de libros digitales.
- Poner al alcance de los profesores, alumnos y la comunidad en general el contenido de los libros que dispone la universidad, para que puedan desarrollar sus actividades educativas y de investigación.

- Desarrollar el sistema de recuperación de la información que se tiene almacenada.
- Promover en las bibliotecas de las facultades de la universidad la implementación de bibliotecas digitales usando estándares (DC, TEI) para preservar y difundir el contenidos de libros digitales así como establecer un espacio en el cual podrán ser consultados.

2.3. Justificación del Proyecto

Debido al constante cambio tecnológico, la Biblioteca Central Pedro Zulen de la UNMSM como parte de la universidad está obligada a mantenerse a la vanguardia, por tal motivo viene incentivando la investigación.

Las bibliotecas digitales hoy en día son una realidad, como toda biblioteca contribuye a la transmisión de conocimiento de generación en generación de manera mas justa y equitativa, y este es un índice que mide el nivel de desarrollo cultural que tiene un país; viendo la importancia del tema , se va investigar metodologías de preservación de datos y los estándares de publicación de material digital tales como XML, TEI, DC y nuevas tecnologías para su desarrollo e implementación.

Para la estructuración, el almacenamiento y ordenamiento de la información se hará de acuerdo a los estándares internacionales de publicación de documentos digitales, todo ello porque permiten la preservación digital a largo plazo, su fácil tratamiento por cualquier sistema de recuperación de información y para el intercambio de conocimiento entre distintas bibliotecas.

Si bien es cierto que en nuestro país somos casi siempre usuarios finales de tecnologías de este tipo, pero también somos capaces de desarrollar los mismos e implantarlos de acuerdo a nuestras necesidades. El prototipo se va implantar en nuestra universidad para contribuir al avance tecnológico y cultural en nuestro país y tener a acceso en línea a los libros de forma gratuita en Internet, de esta manera ampliar su difusión, permitiendo que estos documentos sean un verdadero instrumento de trabajo y de control para satisfacer las demandas de los usuarios finales.

2.4. Alcance

El alcance de este trabajo de tesis está centrada en la preservación y difusión de libros digitales, se ha realizado el prototipo considerando la estructura de un libro estándar sin ninguna selección de algún tipo de libro en especial. En el prototipo se crean los módulos necesarios para el proceso de generación del libro, conversión y recuperación de información almacenada.

2.5. Propuesta de solución

Teniendo en cuenta el poder de la información, en este trabajo de tesis se plantea una metodología para preservar la información, para ello se creará nuevos procesos con el fin de simplificar la generación de documentos digitales, y facilitar el acceso a ellas mediante la biblioteca digital Pedro Zulen de la UNMSM, preservar y poner a disposición de los usuarios.

Implantar una metodología de trabajo para la etapa de la generación de los documentos digitales el cual automatizará el proceso de la generación de documentos XML. HTML.

Capítulo III: Bibliotecas Digitales

3.1. Definición

En estos momentos no es nada fácil definir qué es una biblioteca digital, el término provoca aun una impresión diferente en cada uno de nosotros, para algunos es simplemente la automatización de bibliotecas convencionales y para otros es una colección de manuscritos digitalizados. Para tener una idea mas exacta sobre el concepto, aclararemos las diferencias entre los términos: biblioteca electrónica, biblioteca virtual y biblioteca digital.

“Biblioteca electrónica, es aquella que cuenta con sistemas de automatización que le permiten una ágil y correcta administración de los materiales que resguarda, principalmente en papel. Así mismo, cuenta con sistemas de telecomunicaciones que le permitirán acceder a su información, en formato electrónico, de manera remota o local. Proporciona principalmente catálogos y listas de las colecciones que se encuentran físicamente dentro de un edificio.

Biblioteca virtual, es aquella que hace uso de la realidad virtual para mostrar una interfaz y emular un ambiente que sitúe al usuario dentro de una biblioteca tradicional. Hace uso de la más alta tecnología multimedia y puede guiar al usuario a través de diferentes sistemas para encontrar colecciones en diferentes sitios, conectados a través de sistemas de cómputo y telecomunicaciones

Biblioteca digital, es un repositorio de acervos y contenidos digitalizados, almacenados en diferentes formatos electrónicos por lo que el original en papel, en caso de existir, pierde supremacía. Generalmente, son bibliotecas pequeñas y especializadas, con colecciones limitadas a sólo algunos temas." [66]

En la [web del Digital Library Project], hay una definición de biblioteca digital, que proviene del Santa Fe [Workshop on Distributed Knowledge Work Environments] [27]. Dice así:

"El concepto de biblioteca digital no es únicamente el equivalente de repertorios digitalizados con métodos de gestión de la información. Es más bien, un entorno donde se reúnen colecciones, servicios, y personal que favorece el ciclo completo de la creación, difusión, uso y preservación de los datos, para la información y el conocimiento".

En particular [Gladney et al. 1994] proponen una definición precisa al término y es ampliamente aceptada en el ambiente de la bibliotecas digitales : **" Una Biblioteca digital, es un ensamblaje de cómputo digital, mecanismos de comunicación y software necesario para reproducir, emular y extender los servicio provistos por bibliotecas convencionales como son la colección, catalogación, búsqueda y deseminación de información"**. Un servicio completo de la biblioteca digital, debe estar acompañado de los servicios esenciales de las bibliotecas convencionales además de explorar las ventajas de almacenamiento, búsqueda y comunicación digital. [27]

La Federación de Bibliotecas Digitales en el año 2003 dió una definición que abarca y resume muchas definiciones y expresa de manera clara el concepto de biblioteca digital:

"Las bibliotecas digitales son organizaciones que fomentan los recursos, incluido el personal especializado, para seleccionar, estructurar, ofrecer acceso intelectual, traducir, distribuir, preservar la integridad y garantizar la permanencia de las colecciones digitales, de forma que estén disponibles para una o varias comunidades en un ambiente internacional coordinado." [30]

Las bibliotecas digitales son definidas como repertorios de objetos digitales, más o menos organizados, que sirven a una comunidad de usuarios definida, consideran los derechos de autor presentes y gestionados, y disponen de mecanismos de preservación y conservación. Esta definición tiene en cuenta que estos repertorios constan de datos (el contenido) y metadatos (la información que describe los datos) e incorporan técnicas de búsqueda y recuperación de la información.

“La biblioteca digital no intenta copiar la realidad impresa , sino que genera una nueva estructura de la información que hace que ésta evolucione desde el concepto lineal del libro y los documentos tradicionales al concepto hipertextual, donde la información llega al usuario de formas muy variadas y provista de todo tipo de vínculos, los cuales permiten ampliar, concretar o explicar los contenidos de forma simultánea y diferente. El hipertexto incluye mucha más información no textual que el impreso, ya que incorpora elementos multidimensionales: voz, sonido, imagen, 3 D, etc.”[54].

3.2. Antecedentes y evolución del concepto de bibliotecas digitales

Entre los años sesenta y setenta se desarrollaron pequeños repertorios, o bibliotecas electrónicas, que se basaban sobre todo en la automatización de noticias bibliográficas, algunas veces acompañadas de pequeños resúmenes. Las bibliotecas digitales de esta época incluían únicamente texto. Los primeros ejemplos probablemente se remontan a la automatización de los catálogos de bibliotecas, con sus grandes cantidades de referencias bibliográficas; luego vinieron las bases de datos comerciales, normalmente multitemáticas, que vaciaban publicaciones periódicas, conferencias, etc. de interés para la comunidad investigadora y cuyo formato más habitual era los CD-ROM. Más adelante, con el uso creciente del acceso en línea derivado de los progresos en las comunicaciones y la tecnología informática, los museos, pinacotecas, y otras instituciones culturales decidieron dar un acceso libre a sus fondos a través de estos nuevos medios. Simultáneamente, de forma más o menos espontánea, aparecieron repertorios de documentos de todo tipo generados por diferentes instituciones, grupos e incluso particulares.

En los años noventa, gracias al rápido progreso de las nuevas tecnologías, los repertorios en texto completo aumentaron considerablemente y cambiaron notablemente su visualización, adoptando interfaces cada vez más sencillas de utilizar. También se caracterizaron por incorporar nuevos elementos: texto, datos, imágenes, figuras 3D, gráficos, vídeos, audio, etc. y nuevas tecnologías: gestores de bases de datos, sistemas de información geográfica, hipertexto, sistemas multimedia, lenguaje natural, procesamiento y recuperación de la información, etc.

En Estados Unidos el interés por las bibliotecas digitales está creciendo gracias a la creación en 1994 de la Digital Libraries Initiative, patrocinada

por la NSF (National Science Foundation), la DARPA (Defense Advanced Research Projects Agency) y la NASA. La misión de este organismo es impulsar el progreso tecnológico en la recolección, almacenamiento y organización de la información digital y en los métodos que permitan hacerla accesible para la búsqueda, recuperación y procesamiento de los datos mediante las redes de comunicaciones. Estos desarrollos tienen un gran interés para la educación, sobre todo la educación a distancia (universidades virtuales), así como para la difusión de información por ejemplo, en las bibliotecas y los museos virtuales; también porque permiten el tratamiento de todo tipo de materiales, imágenes, audio, vídeo, etc.

Las bibliotecas digitales nacen gracias al avance tecnológico vertiginoso de las últimas décadas, especialmente con la aparición de las redes informáticas e internet como medio de publicación e interconexión, la existencia del profesional informático, la existencia de los libros físicos y la informática en sí como una herramienta de apoyo, estos son los elementos claves para la existencia de las bibliotecas digitales, como vemos en la [Figura 1], son los pilares del desarrollo siendo necesarios y dependientes todos los elementos mencionados.

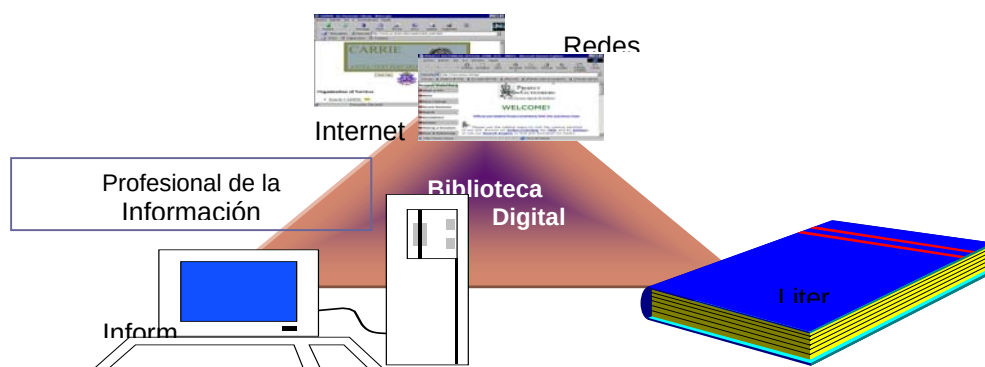


Figura 1: Elementos de la biblioteca digital

Uno de los campos en que están muy desarrolladas las bibliotecas digitales es el ámbito de las ciencias biomédicas. El tratamiento digital de las imágenes médicas en Estados Unidos (resonancias magnéticas, angiografías, tomografías, etc.) representa un 30% de los reconocimientos en imágenes médicas. El resto son rayos X y radiografías. Entre las bibliotecas más destacadas actualmente están las siguientes bibliotecas:

- Biblioteca de la OMS

(<http://www.who.int/library/index.es.shtml>), sus servicios de información están destinados al sector de la salud, los programas de la Biblioteca de la OMS ayudan a las regiones y a los países en desarrollo a ser autosuficientes

- Biblioteca de vídeos médicos

(<http://www.crd.ge.com/esl/cgsp/projects/video/medical/index.html>) está especializado en el manejo de imágenes y videos en 2D y 3D del campo médico.

- Biblioteca digital dioscorides. Anatomía, balnearios, cirugía y farmacología ubicada en la siguiente url:

(http://cisne.sim.ucm.es/search*spi~S4/e?SEARCH=biblioteca+digital) Esta biblioteca digital incluye 2000 libros del fondo antiguo digitalizados a texto completo entre los que se pueden consultar, otros; además de estos temas relacionados con las Ciencias de la Salud. El proyecto ha sido llevado a cabo por la Biblioteca de la Universidad Complutense, la Fundación Ciencias de la Salud y los Laboratorios GlaxoSmithKline .

- Biblioteca Médica Digital (Argentina) ubicada en la siguiente url: (<http://www.bmd.com.ar/>) tiene como objetivo principal llevar la medicina por especialidades y patología, a ser

un material de consulta multimedia completo y permanente de fácil manejo, con ahorro de tiempo, aprovechando el tratamiento estándar de esta tecnología y su amplia difusión.

- Biblioteca Nacional Alemana de Medicina, ubicada en la siguiente url: <http://www.zbmed.de/index.html>, su contenido está relacionado con material de la ciencia médica.

- Biblioteca Nacional de Medicina-USA, ubicada en la siguiente url: <http://www.nlm.nih.gov/>, es una de las bibliotecas más completas, tiene todos los materiales necesarios como revistas, libros, proyectos médicos ,etc.

- Biblioteca Virtual de CSIC, ubicada en la siguiente url: <http://www.csic.es/cbic/cbic.htm>, es una biblioteca que viene a ser el punto único de acceso a los recursos electrónicos del CSIC (bases de datos, revistas y libros electrónicos, portales, catálogos,etc.) que pone a disposición de los usuarios una amplia gama de servicios para la gestión de la información: búsquedas simultáneas, navegación entre recursos, acceso a texto completo, fondos en las bibliotecas del CSIC, solicitud de fotocopias y préstamos, descarga de registros, factores de impacto, etc.

- Bireme(Brasil), ubicada en la siguiente url: <http://www.bireme.br/>, una de las bibliotecas más importante para el campo médico, tiene un amplio contenido y enlaces a otras bibliotecas de este rubro, entre sus contenidos tienen búsqueda en bases de datos, pedidos de fotocopias - SCAD, portales de revistas científicas SciELO.org, localizador de información en salud.

- Scielo (La Scientific Electronic Library Online), ubicada en la siguiente url: <http://www.scielo.org.pe/scielo.php>, Es una biblioteca virtual que abarca una colección seleccionada de revistas científicas brasileñas y de América latina.
- Consorcio de bibliotecas relacionadas con salud, ubicada en la siguiente url: <http://www.hslc.org/>.
- Instituto Karolinska - Suecia , ubicada en la siguiente url: <http://ki.se/ki/jsp/polopoly.jsp?d=167&l=en>, es una biblioteca dirigido al campo médico que contiene catálogos, lista de revistas, libros electrónicos más de 20 000 publicaciones médicas, muchas bases de datos, diccionarios y enciclopedias.
- Library of the National Medical Society, ubicada en la siguiente url: <http://www.ccspublishing.com/index.htm>, es una de las bibliotecas de la sociedad médica que contiene muchos artículos, libros y revistas especializadas por cada area médica, diagnósticos, libros de estrategias clínicas, etc.

3.3. Características

Las bibliotecas digitales, en la medida en que dan acceso a grandes repertorios de información, tienen que desarrollar herramientas y tecnologías para dar valor añadido al acceso del conocimiento y al significado inherente a los contenidos de las colecciones digitales. La rápida expansión de Internet y el uso generalizado de la World Wide Web han producido una evolución y revolución en los productos documentales ofertados por los editores. Las bases de datos en CD-ROM están siendo sustituidas cada vez más por el acceso en línea, aunque la velocidad de las conexiones todavía no está, en algunos casos, a la altura

de lo que sería aceptable, con las ventajas que ello supone para la actualización inmediata de los datos y la no importancia de la ubicación, todo lo cual hace que el producto sea mucho más atractivo.

El valor de las bibliotecas digitales depende de la calidad de los contenidos y de su organización, así como de los sistemas de gestión que facilitan el acceso y visualización de los datos almacenados. Es muy importante que haya colaboración entre informáticos, profesores, bibliotecarios y documentalistas en el diseño de las bibliotecas digitales. También es importante un sistema de comprobación y seguimiento continuo que permita ver la evolución del diseño adoptado y recoja las respuestas de los usuarios.

La Association of Research Libraries (ARL) [29] señala unos elementos comunes a los diversos términos con los que se designan las bibliotecas digitales (bibliotecas electrónicas, bibliotecas virtuales, etc.). Algunos de estos elementos son:

- La biblioteca digital no debe ser una entidad individual.
- La biblioteca digital requiere que haya medios tecnológicos para enlazar recursos.
- Los enlaces entre un gran número de bibliotecas digitales y los servicios de información deben ser transparentes para los usuarios.
- El acceso universal a las bibliotecas digitales y a los servicios de información debe ser un objetivo principal.
- Las bibliotecas digitales no deben limitarse a suplir documentos, sino que deben ofrecer otros elementos digitales que no pueden suministrarse en formato impreso.

Una de las características de las bibliotecas digitales, es que la información que contienen ha sido creada por gente diversa, utilizando medios diversos, dándole formas y formatos diferentes, almacenada en diferentes lugares del mundo (servidores) y de manera creciente e interconectada por medio de redes. Es decir, en estas bibliotecas conviven materiales en diferentes formatos, en distintas versiones, ubicados en diferentes lugares, y accesibles a un gran número y diversidad de personas.

Los proyectos de bibliotecas digitales y la investigación en estos temas deben permitir el cambio continuo, debido al aumento del ancho de banda de las redes de comunicaciones, las cuales permiten gestionar y dar coherencia, utilizar y posibilitar el acceso a gran cantidad de datos distribuidos y transformados en información y conocimiento.

La existencia de las bibliotecas digitales hace cada vez más necesario que haya sistemas de recuperación de información que sean capaces de procesar el lenguaje natural. "Estos sistemas recuperan y seleccionan frases lingüísticas como unidades de información y términos controlados que forman parte de tesoro, o términos incluidos en una estructura de árbol del conocimiento.

Estos sistemas de recuperación tienen que ser:

- **Flexibles:** capaces de procesar diferentes tipos de información
- **Precisos:** capaces de seleccionar información pertinente y desestimar el "ruido".
- **Rápidos:** tiene que poder tratar simultáneamente cantidades ingentes de información y documentación
- **Automáticos:** capaces de seleccionar la información sin que tenga que estructurarse antes

- **Fáciles:** su utilización no tiene que suponer un problema para el usuario “[29]

Concretaremos las características de las bibliotecas digitales desde diferentes ámbitos:

3.3.1. Contenidos

Las características de estos textos son diferentes de las habituales en los textos impresos. La estructura utilizada por los escritores se vuelve diferente. El escritor W.I. Arms afirma que él escribe de una forma cuando lo hace, por ejemplo, en el D-Lib Magazine, revista en línea accesible por Internet, y de otra cuando escribe para una revista de formato tradicional.

Las ventajas de la digitalización se concretan en características como: facilidad de acceso, velocidad en la recuperación de la información, transmisión en línea de forma remota, gran capacidad de almacenamiento, etc. En el capítulo de contenidos cabe destacar que a la información textual se añade información no textual (fotografías, dibujos, ilustraciones, etc.), gran cantidad de datos numéricos (información proveniente de satélites, datos cosmológicos, etc.), sonidos digitalizados, imágenes en movimiento, representaciones multidimensionales (hologramas), entre otros elementos.

Tan importante como la inclusión de estos tipos de contenidos son sus características como elementos abiertos, públicos, dinámicos, en cambio constante, las cuales permiten la colaboración simultánea de diferentes personas en un mismo trabajo.

La organización de los materiales que configuran las bibliotecas digitales, así como el diseño y la estructura que presentan, están muy condicionadas por las tecnologías que las hacen posibles y por las redes de comunicaciones que facilitan el acceso de estos contenidos.

3.3.2. Preservación

La introducción de la tecnología digital implica el desarrollo de infraestructuras para hacer accesibles los materiales que quieren preservarse.

Respecto a esto, cabe mencionar que en Estados Unidos existe el Digital Preservation Consortium (DPC)[23]. A este organismo lo forma un grupo de universidades del mundo que trabajan conjuntamente, tanto a nivel de investigadores como de estudiantes para conseguir un uso efectivo de los materiales preservados en formato digital. La misión de este consorcio es desarrollar el uso y la utilidad de la tecnología digital para preservar e implementar el acceso a obras intelectuales de importancia nacional e internacional.

Las universidades miembros de este consorcio son: la University of California, Berkeley, la Columbia University, la Cornell University, la Harvard University, la University of Michigan, la Pennsylvania State University, la Princeton University, la Stanford University, la University of Southern California, la University of Tennessee y la Yale University.

Éstos son los objetivos que se han propuesto para los próximos 5-10 años:

- Verificar y monitorizar el uso de las imágenes digitales para su preservación y acceso.

- Definir y promover métodos y criterios compartidos, para la producción, almacenaje y distribución de imágenes digitales .
- Ampliar la base de materiales preservados en forma de imágenes digitales .
- Desarrollar y mantener mecanismos fiables que garanticen el acceso generalizado a documentos digitales de imágenes .

3.3.3. Flexibilidad

La presencia de elementos multimedia en estos repertorios hace que tengan que integrarse diferentes tipos de tecnologías básicas en su construcción. En consecuencia, tienen que ser elementos flexibles que permitan estas integraciones y que garanticen la adaptación a los constantes cambios provocados por la rápida evolución de los tipos de contenido.

3.3.4. Cooperación

El éxito de las bibliotecas digitales estará muy condicionado por la cooperación. El trabajo interdisciplinario cada vez más será un elemento fundamental en este contexto. El crecimiento exponencial de los datos hará necesaria una coordinación a gran escala. El trabajo conjunto de científicos, ingenieros, informáticos, editores, bibliotecarios, etc. será esencial para crear y adaptar nuevas tecnologías y desarrollar nuevas normativas.

3. 3.5. Tecnología

Teniendo en cuenta las características de las bibliotecas digitales, serán elementos muy importantes que la tecnología usada en ellas cumplan con

las siguientes características: la interoperabilidad, la adaptabilidad, la rapidez, la estandarización, y la seguridad.

3.3.6. Legislación

Los derechos de autor y la legislación sobre la propiedad intelectual son aspectos de vital importancia tanto para la creación de estas bibliotecas como para su protección. La preocupación por este tema, es muy presente en todos los ámbitos relacionados con las bibliotecas digitales en Estados Unidos; en el ámbito europeo se ha promovido la promulgación de la Directiva 96/9/CE del Parlamento Europeo y del Consejo del 11 de marzo de 1996, sobre la protección jurídica de las bases de datos. Estados Unidos y la Unión Europea intentan llegar a un tratado multilateral sobre la protección de las bases de datos.

En nuestro país aun no esta contemplado estos derechos y mas bien tenemos la “Ley de democratización del libro y de fomento de la lectura” que mas bien promueve la publicación de los libros tanto en formato de papel y digital, pero en cuanto al tema digital dice textualmente lo siguiente: “La reproducción por medios reprográficos, digitales u otros, creados o por crearse, de libros y productos editoriales afines, se registrá por lo establecido en la legislación sobre derechos de autor y la Decisión 351 del Acuerdo de Cartagena.” [24]. Para mayor detalle del acuerdo de cartagena ver el [Anexo 1].

3.3.7. Formatos

El crecimiento exponencial de la información en línea en estos últimos años ha traído consigo un crecimiento paralelo de medios y formatos. Cualquier usuario puede crear un documento en un determinado formato, almacenarlo en un formato diferente al original, y/o enviarlo

por correo con cualquier forma creada mediante software. Por lo común, este formato está definido por las especificaciones elaboradas por una empresa determinada. En algunos casos la definición puede ser producto de unos estándares formales. Los diferentes formatos pueden competir o complementarse. Los formatos utilizados en estos momentos para la confección de bibliotecas digitales pueden ser de tipo abierto o cerrado: Son formatos de tipo abierto, es decir, pueden manipularse con más facilidad, los formatos derivados del SGML: HTML, XML, etc. Son formatos de tipo cerrado, es decir, presentan más dificultades para ser cargados, modificados o manipulados, como el Acrobat, PDF, etc.

Algunas instituciones que han implementado bibliotecas digitales están creando sus propios formatos, por ejemplo, en el caso de la University of California (Berkeley) donde han desarrollado el formato MVD (multivalent documents). El MVD es un nuevo modelo de documento altamente extensible, compatible con múltiples plataformas, soporta documentos en diferentes formatos y permite una composición y distribución de los datos y de programas coherentes.

La facilidad de empleo de la documentación digital (transferencia de copias, cambios de soporte, cambios de formato, etc.) hace que los editores y otros colectivos afines alcen sus voces para reclamar que se tengan en cuenta los derechos de autor.

3.4. Metadato

Paralelamente al gran desarrollo de las bibliotecas digitales ha surgido la necesidad de procesar los contenidos de estos repertorios para facilitar la búsqueda y la recuperación de la información de una forma eficaz.

3.4.1. Antecedentes y definición

“Los metadatos han estado con nosotros desde que el primer bibliotecario hizo una lista de los documentos que tenía en un estante, en un pergamino escrito a mano. El término "meta" viene de una palabra griega que significa "junto a, con, después, siguiente". El uso más reciente del latín y el inglés emplearía "meta" para indicar algo trascendental, o sobrenatural. Los metadatos, pues, pueden ser definidos como datos sobre otros datos. Es el término usado en la era de Internet para la información que los bibliotecarios tradicionalmente habían puesto en los catálogos, y más comúnmente se refiere a información descriptiva sobre recursos de la Web. Un registro de metadatos consiste en un conjunto de atributos, o elementos, necesario para describir la fuente en cuestión. Por ejemplo, un sistema de metadatos común entre los bibliotecarios (el catálogo de biblioteca) contiene un conjunto de registros de metadatos con elementos que describen un libro u otra publicación en una biblioteca: Autor, título, fecha de creación o publicación, materia, y la signatura topográfica especificando la localización de la publicación en el estante.”[\[40\]](#).

Según Howe (1993) [\[45\]](#), el término dato sobre el dato fue acuñado por Jack Myers en la década de los 60 para describir conjuntos de datos. La primera acepción que se le dió (y es actualmente la más extendida), ya que proporcionaban la información mínima necesaria para identificar un recurso. En este mismo trabajo se afirma que puede incluir información descriptiva sobre el contexto, calidad y condición o características del dato.

La evolución del término desde esta fecha hasta 1997 ha sido descrita por Lange y Winkler (1997) revelando que no existen demasiadas novedades.

Atendiendo a la definición antes mencionada, podríamos considerar la catalogación como un proceso de generación de metadatos. Teniendo en cuenta que la mayoría de sistemas de metadatos ha sido creada no sólo por profesionales de la información sino también por informáticos, diseñadores de programas, técnicos de sistemas, etc., la utilización de este término puede conllevar una carga excesiva (por ejemplo, reglas de catalogación, clasificaciones de materias, etc.).

“La mayoría de funciones descritas por los autores sobre metadato las podemos encontrar agrupadas en el trabajo de Iannella y Waugh (1997):

- Resumir el significado de los datos
- Permitir la búsqueda
- Determinar si el dato es el que se necesita
- Prevenir ciertos usos (PICS)
- Recuperar y usar una copia del dato
- Mostrar instrucciones de cómo interpretar un dato
- Obtener información sobre las condiciones de uso (derechos de autor)
- Aportar información acerca de la vida del dato
- Ofrecer información relativa al propietario/creador
- Indicar relaciones con otros recursos
- Controlar la gestión

El padre del Web, Tim Berners-Lee, se percató rápidamente de la importancia de los metadatos. Para él, su concepto no debía limitarse a la descripción de recursos Web. Más bien se debía ampliar, englobando las particularidades de gente, cosas, conceptos e ideas (Berners-Lee, 1997). Si bien es cierto que su definición es ambiciosa, Berners-Lee no contempló la posibilidad de extrapolar sistemas de metadatos a otros recursos

electrónicos que no fueran Web. Para Berners-Lee existen tres tipos de metadatos en el Web:

- El primero de ellos es el que se encuentra dentro del documento mismo (por ejemplo aquellos que se pueden encontrar en cualquier documento generado por un procesador de textos).
- El segundo es el que se produce durante una transferencia HTTP (HyperText Transfer Protocol) cliente y servidor se envían información sobre el objeto que están transmitiendo por medio de metadatos.
- El último es más difícil de encontrar, ya que el metadato se utiliza cuando se consulta en otro documento (para comprobar si se puede acceder a él -o al sitio Web-, verificar derechos de autor...).”[\[41\]](#) .

La relación entre un registro de metadatos y el recurso al que describe en resumen puede darse de una de estas dos formas:

1. Los elementos pueden estar en un registro separado del documento, como en el caso del registro de un catálogo de bibliotecas; o
2. los metadatos pueden estar incluido en el propio recurso.

Ejemplos de metadatos incluidos al propio recurso son, por ejemplo, los datos de la Catalogación en Publicación impresa en el verso de la página del título de un libro, o la cabecera TEI en un texto electrónico. Muchos estándares de metadatos utilizados hoy en día, incluido el Dublin Core, no prescriben ningún tipo de relación, dejando así la decisión a cada caso de implementación particular.

“A pesar de que el concepto de metadatos antecede a Internet y a la Web, el interés mundial por las normas y prácticas de metadatos ha estallado con el crecimiento de la publicación electrónica y las bibliotecas digitales, y la concurrente "sobrecarga de información" que resulta de las grandes cantidades de datos digitales disponibles en línea. Cualquiera que pretenda encontrar información en línea utilizando uno de los servicios de búsqueda Web utilizados hoy en día, probablemente habrá experimentado la frustración al encontrar cientos, o miles, de ocurrencias con la posibilidad limitada de refinar o de hacer una búsqueda más precisa. La adopción a gran escala de estándares y prácticas descriptivas para los recursos electrónicos mejorará la recuperación de recursos relevantes en cualquier contexto donde la recuperación es crítica. Como señalan Weibel y Lagonze, dos líderes en el campo del desarrollo de metadatos.” [\[40\]](#)

La biblioteca digital tiene que cumplir una serie de características que le den el valor que necesita para difundir estos contenidos. Tiene que ser recuperables mediante metadatos (datos de los datos) que proporcionen valor añadido a la mera acumulación de información. Los metadatos tienen gran importancia en la composición de las bibliotecas digitales, ya que permiten búsqueda efectiva y precisa. En este campo, hay que destacar los trabajos normativos desarrollados por Dublin Core .

3.4.2. Convenciones, sistemas de metadatos

Son las convenciones más extendidas y satisfacen los siguientes requerimientos son : [\[41\]](#)

- Identificación de documentos en un entorno distribuido.
- Descripción de su contenido.
- Localización y accesibilidad.

- Gestión de derechos: copyright, reproducción, restricciones de acceso.

Las convenciones que más destacan y que son aceptados por la norma HTML, se detallan a continuación :

- **DC o DCMI** (Dublin Core Metadata Initiative).- Nacido como un punto intermedio entre MARC y TEI. Recursos electrónicos. Ambito educativo. Rss, Ebooks y otros.
- **RDF** (Resource Description Framework).
- **TEI** (Text Encoding Initiative).-Material textual. Originalmente usado por investigadores, adoptado por bibliotecarios.
- **MARC DTD** (Machine Readable Cataloging Document Type Definition).
- **EAD** (Encoded Archival Description).- Descripción de documentos de archivos.
- **PICS** (Platform for Internet Content Selection).- Para permitir o no el acceso a determinados contenidos
- **MCF** (Meta Content Format).- Desarrollado por Apple, describe cómo y qué contenido se almacena.
- **IAFA** (Internet Anonymous FTP Archive).- Como ficheros de intercambio en sistemas de indización distribuida
- **SOIF** (Summary Object Interchange Format).- Como ficheros de intercambio en sistemas de indización distribuida
- **LOM** (learning Objects Metadata).- Para la descripción de objetos individuales aplicables a educación a distancia (9 cate 47 elem).
- **GILS**.-USA relacionada con la ubicación de la información producida por gubernamentales.
- **LDAP-DIF** .- Para intercambio de información entre LDAP's
- **Music-Brainz** .- Música. Basado en XML/RDF

- **NBII** .- Perfil biológico de metadato, derivado del estándar FGDC . Para información biológica. Complejo. Administración de la clasificación taxonómica.
- Y muchos otros...Hasta un conjunto SGML para crear metadatos hasta de armas nucleares. Se encuentra especificada en la ruta: (<http://www.osti.gov/inforum98/mason.html>).

La mayoría de estos sistemas se utilizan de manera aislada ya que han sido creadas con requerimientos muy específicos.

3.5. Problemática de bibliotecas digitales

Estas son la problemáticas actuales según Dora Pérez [53], en donde nos explica que, muchos autores han presentado al Internet como una gran biblioteca digital. No podemos considerar que un gran almacén de contenidos diversos, presentado en diferentes formatos, estructurado de diferentes maneras, organizado según los diferentes intereses que los han originado y que persiguen finalidades variadas, pueda considerarse una biblioteca digital.

La avalancha de información existente en la red pone de manifiesto la importancia del profesional que tiene que guiar al usuario en la búsqueda y selección de la información conseguida.

Segun Dora Pérez [53] para construir un gran sistema de bibliotecas digitales, hay en estos momentos tres áreas problemáticas y que no tienen una fácil solución, si no es con la ayuda del trabajo interdisciplinario:

- Los cambios tecnológicos
- los cambios económicos
- los cambios sociales y culturales

3.5.1. Cambios tecnológicos

La búsqueda y el uso de la información conocen hoy día una rápida transformación. El progreso tecnológico, el cambio que se está produciendo en la velocidad y el ensanchamiento de banda de las comunicaciones, el uso de protocolos de acceso abierto, como la web, entre otros aspectos, abren toda una nueva perspectiva amplia y diversa.

A lo largo de estos últimos años se han ido desarrollando gran número de programas que permiten la consulta, la búsqueda y la recuperación de la información contenida en diferentes bibliotecas digitales o bases de datos de contenidos diversos. A pesar de eso, aunque se han ido solucionando problemas relacionados con estos elementos, todavía hay cuestiones técnicas pendientes, tales como:

- Cómo se comunican las bibliotecas entre sí.
- Cómo tiene que buscar el usuario determinada información.
- Cómo deben clasificarse, seleccionarse, combinarse y entregarse los resultados de la búsqueda al usuario.
- Cómo se asegura la integridad y la seguridad de los datos.

En este sentido, las bibliotecas digitales requieren un correcto cumplimiento de la legislación sobre los derechos de autor, en lo relativo a la gestión de estos derechos por un mayor número de usuarios y en una diversidad de formatos. Asimismo, tienen que asegurar que esos datos no

se manipularán ni difundirán sin autorización, así como garantizar al usuario que la información no ha sido alterada.

3.5.2. Cambios económicos

Los costos de las bibliotecas digitales afectan a tres niveles:

- Infraestructuras de comunicaciones
- Infraestructuras de equipamiento (por ejemplo los ordenadores)
- Contenidos (las bases de datos en las cuales se van almacenar los libros)

3.5.3. Cambios sociales y culturales

Hasta ahora, se aseguraba el acceso y consulta de los fondos depositados en las biblioteca tradicionales, tanto públicas como universitarias, de forma libre y gratuita.

¿Ocurrirá lo mismo con las bibliotecas digitales?, ¿podrá asegurarse el acceso gratuito a los contenidos?. ¿en qué casos?, ¿y el acceso a los medios tecnológicos que permiten ese acceso?

Pese a que es evidente que el acceso a las redes cada vez está más extendido, es preciso tener en cuenta que todavía dista mucho de ser un elemento común a todos los estamentos sociales.

Kuny y Cleveland [\[25\]](#) señalan una serie de aspectos negativos en la generación de estas bibliotecas. Estos autores intentan desmitificar la opinión que la mayoría de las personas tiene sobre la biblioteca digital, y hacen unas reflexiones sobre los peligros, inconvenientes o problemas que éstos nuevos desarrollos pueden presentar, poniendo de manifiesto las

limitaciones tecnológicas, las limitaciones legales, y los obstáculos administrativos. Uno de los mitos que intentan combatir es la opinión de que la biblioteca digital es sinónimo de Internet.

También tienen en cuenta el costo económico de los proyectos de digitalización. En estos momentos los costos inherentes a los procesos de digitalización no son todavía ni rápidos, ni fiables al cien por ciento, ni mucho menos económicos.

En el caso de bibliotecas digitales que contengan elementos multimedia, se pone de manifiesto que los requerimientos de hardware y software para contener y acceder a estos materiales no está todavía al alcance de todo el mundo, ni siquiera de una mayoría de usuarios de las redes, ya sea por razones técnicas o económicas.

En estos momentos las principales fuentes de información organizadas y con contenidos de interés, sobre todo para la investigación, a las que podemos acceder son de pago, es decir, no son de libre acceso y consulta si no es mediante una cuota, ya sea en forma de suscripción o en forma de pagar por ver (pay per view).

Dos conocidos autores, que han tenido un importante papel en el campo de la automatización de bibliotecas, Walt Crawford y Michel Gorman [\[6\]](#), en una obra cuyo subtítulo es muy significativo, Sueños, locuras y realidades, en el cual ponen de manifiesto la problemática y la realidad de algunas afirmaciones que se están generalizando sobre las bibliotecas digitales. Estos autores intentan demostrar que no puede hablarse de bibliotecas digitales como bibliotecas que sustituirán a las actuales bibliotecas tradicionales, que hay muchas premisas dadas como ciertas que no lo son, o que sólo lo son en parte, y que aun estamos muy lejos de

las auténticas bibliotecas digitales y de todas las ventajas que se les atribuyen.

Dichos autores concretan unas características que los puntos de trabajo de los usuarios virtuales (herramienta esencial para acceder a las llamadas bibliotecas digitales) deberían ofrecer para que tuvieran sentido para los mismos usuarios y para la sociedad, y llegan a la conclusión de que en estos momentos no se cumplen. Estas características son:

- **Disponibilidad:** todo lo que existe registrado (impreso, fotografiado, filmado, pintado, dibujado, etc.) tendría que convertirse a formato digital a fin de que éste disponible para todos los usuarios con un terminal de trabajo.
- **Recuperación y adecuación:** cada usuario de este hipotético terminal de trabajo (que permitiría el acceso a la biblioteca digital) tendría que poder acceder a todos los documentos electrónicos relevantes de este universo digital, de una manera rápida y fácil.
- **Autenticidad:** cada usuario debería tener la seguridad de que el documento que encuentra en la red es el documento auténtico y original.
- **Utilización:** cada uno de los documentos recuperados mediante el terminal de trabajo, tendrían que ser recuperados de forma que todo usuario pudiera usarlo.
- **Protección de la propiedad intelectual:** la protección de los derechos de autor debería estar garantizada en todo documento recuperado.
- **Asequibilidad:** Los costes de acceso y recuperación de los diversos documentos tendrían que ser razonables y no superar los costes de sus equivalentes tradicionales.

Según los especialistas estas características no se cumplen en estos momentos y no se sabe si serán posibles en un futuro próximo. Dichos estudiosos creen que, hasta que no se consigan estas características no habrá un terminal de trabajo universal, y sin ello no será posible un futuro donde todo sea electrónico. En realidad creen que una biblioteca digital nunca podrá sustituir al medio impreso y sus análogos.

Capítulo IV: Preservación digital

4.1. Introducción

“La importancia de los documentos digitales como medio de transmitir la información y el conocimiento es cada día mayor tanto cuantitativa como cualitativamente. La consulta de los objetos digitales, que pueden ser textos, datos, gráficos, imágenes fijas o en movimiento, grabaciones sonoras, programas informáticos, etc., está al alcance de amplios sectores de la población y les permite acceder a todo tipo de información. Los documentos de esta clase se generan directamente en formato digital o se convierten a éste a partir de material analógico ya existente.

Sin embargo, pese a su difusión universal y a que constituyen un medio de expresión insustituible del conocimiento y la creatividad humana, los documentos digitales tienen unas características físicas que los hacen mucho más efímeros, volátiles y vulnerables que los documentos tradicionales en papel. De hecho muchos documentos de origen digital, que sólo existían en formato electrónico, han desaparecido ya y son irrecuperables.

La preservación del Patrimonio Digital o Preservación Digital es una preocupación presente en los actuales planes de actuación de los principales organismos internacionales relacionados con el mundo de las bibliotecas y del Patrimonio Cultural. Prueba de este interés internacional

por la conservación de los recursos electrónicos son los documentos básicos sobre el particular aparecidos en los últimos años:

- o *Carta para la preservación del Patrimonio Digital*. UNESCO, 2003.
- o *Resolución del Consejo de 25 de junio de 2002 sobre «Conservar la memoria del mañana: Conservar los contenidos digitales para las generaciones futuras»* (2002/C 162/02). Consejo de la Unión Europea , 2002.
- o *Preserving the Memory of the World in Perpetuity: A Joint Statement on the Archiving and Preserving of Digital Information*. IFLA and IPA (International Publishers Association), 2002.
- o *i2010: Bibliotecas digitales. Comunicación de la Comisión al Parlamento Europeo, al Consejo. Al Comité Económico y Social Europeo y al Comité de las Regiones*. Bruselas, 2005.

En marzo de 2003 la División de la Sociedad de la Información de la UNESCO publicó *Directrices para la preservación del Patrimonio Digital* , preparadas por la Biblioteca Nacional de Australia, una de las instituciones más preocupadas por esta materia. “[67]

“Historia: El problema de la fragilidad de los datos electrónicos emerge con fuerza en el ámbito de las bibliotecas universitarias a mediados de la década de los 90. Entre las primeras llamadas a la acción se encuentra un artículo de Jeff Rothenberg, publicado en la revista *Scientific American [Investigación y Ciencia]* de marzo de 1995, cuyo título pregunta: “¿Son perdurables los documentos digitales?”. El autor plantea la situación hipotética de dejar un CD-ROM con instrucciones sobre cómo encontrar su fortuna y especula con los problemas que tendrían sus nietos en intentar descifrar el texto. Concluye el artículo, contestando la pregunta planteada en el título: si el CD incluye, además de la clave, “toda la base informática requerida juntamente con especificaciones completas y fáciles

de descifrar concernientes al equipo físico necesario, tal vez... [se logre] generar un emulador capaz de ejecutar los programas originales encargados de mostrar“ el documento. Aunque escrito en tono jocoso, subraya un problema muy serio. Escrito ya hace 8 años, se puede afirmar que una solución definitiva aun nos elude. De todas formas no es por una falta de acción por parte de las partes interesadas, que han estado trabajando con diligencia para encontrar soluciones. El problema es debido a la enorme complejidad más allá de lo que pueda parecer a primera vista, porque por debajo de los problemas técnicos que se pueden ir resolviendo, se descubren capas más difíciles de roer, como por ejemplo aspectos administrativos, económicos y legales.

A partir del artículo de Rothenberg, se crearon varios frentes en el mundo anglosajón:

Casi de forma simultánea, en 1996 se elaboraron dos informes en EE.UU. y en Gran Bretaña: En EE.UU. se publicó el informe *Preserving digital information*, subvencionado por el Research Libraries Group (RLG) (una cooperativa de las grandes bibliotecas de investigación, con presencia mayoritaria de las universitarias) y la Comisión on Preservation and Access (CPA que luego fue incorporado en el Council on Library and Information Resources CLIR). Este documento representa el primer paso hacia la definición del problema y proponía recomendaciones para dar una respuesta al problema. Una de ellas fue para la realización de un estudio que cuantificase la verdadera situación en las bibliotecas. Este se llevó a cabo en 1998 y los resultados se publicaron en el documento de M. Hedstrom, *Digital preservation needs and requirements in RLG member institutions*.

Otro documento pionero clave fue el informe de Paul Conway, director de preservación en la biblioteca de la Yale University, publicado también

en 1996 por la CPA, en que se preveían los cambios que iban a ser necesarios para adaptar las operaciones tradicionales de preservación de materiales analógicos y enfrentarse al reto de los materiales digitales.

Mientras, al otro lado del Atlántico el Joint Information Services Committee (JISC) junto a la British Library, organizaron unas jornadas sobre la preservación a largo plazo de los materiales electrónicos, cuyo informe se publicó en 1996, bajo la dirección de M.Fresko. Este primer esfuerzo sirvió de guía para numerosos proyectos posteriores financiados por la administración británica. “[72]

En nuestro país no se dispone de un plan de conservación del Patrimonio Digital similar a los que existen en otros países, “como PANDORA (Australia), DNEP (Holanda), KULTURAW3 (Suecia), NDIIPP (EE.UU.) o NEDLIB (varios países Europeos), desde hace varios años. En estas circunstancias podemos sospechar que gran parte de los recursos electrónicos publicados en Internet en los últimos años del siglo XX, los incunables de la edición electrónica, no están disponibles ni se pueden consultar en nuestros días.”[67]

4.2. Objetivo

“Mantener la capacidad de visualizar, recuperar , manipular y aprovechar la información digital pese a los constantes cambios de los dispositivos electrónicos.”[68]

“Los asuntos que se deben tratar en la preservación digital incluyen lo siguiente:

- Mantener la fiabilidad física de los archivos, los metadatos complementarios, textos y programas (por ejemplo: asegurarse de que el medio de almacenamiento es confiable, con copias de

seguridad (back-ups), mantener la infraestructura de hardware y software necesaria para almacenar y proporcionar acceso a la colección);

- Asegurar el uso de la colección de imágenes digitales en forma continuada (por ejemplo: mantener una interfase de usuario actualizada, permitir a los usuarios recuperar y manipular información para poder satisfacer sus necesidades de información);
- Mantener la seguridad de la colección (por ejemplo: implementar estrategias para controlar la alteración no autorizada de la colección, desarrollar y mantener un programa de gestión de derechos para servicios con cargo)”[70].

4.3. Herencia y Patrominio Digital

“En los documentos de la UNESCO, el patrimonio se define como nuestra herencia del pasado, nuestros bienes actuales y lo que legamos a las generaciones futuras. El patrimonio es, o debería ser, algo que se transmite de generación en generación porque se valora.

La noción de patrimonio cultural es conocida: se trata de aquellos lugares y objetos tangibles e intangibles que poseen valor cultural, histórico, estético, arqueológico, científico, etnológico o antropológico para determinados grupos o individuos. También el concepto de patrimonio natural es muy familiar: características físicas, biológicas o geológicas, hábitats de especies de la flora o la fauna y regiones que poseen valor por razones científicas o estéticas, o desde el punto de vista de la conservación.

¿Existe realmente un patrimonio digital?

Según el Proyecto de Carta para la Preservación del Patrimonio Digital:

Cada vez más, los recursos que son fruto del saber o la expresión de los seres humanos, sean éstos de carácter cultural, educativo, científico y administrativo, o la información técnica, jurídica, médica y de otras clases, se generan directamente en formato digital o se convierten a éste a partir de material analógico ya existente. Los productos “de origen digital” no existen en otro formato que el electrónico original.

Los objetos digitales pueden ser textos, bases de datos, imágenes fijas o en movimiento, grabaciones sonoras, material gráfico, programas informáticos o páginas Web, entre otros muchos formatos posibles dentro de un vasto repertorio de diversidad creciente. A menudo son efímeros y su conservación requiere un trabajo específico en este sentido en los procesos de producción, mantenimiento y gestión. Muchos de esos recursos revisten valor e importancia duraderos y constituyen por ello un patrimonio digno de protección y conservación en beneficio de las generaciones actuales y futuras. Este legado puede existir en cualquier lengua, en cualquier parte del mundo y en cualquier campo de la expresión o el saber humano.

Gracias a las computadoras y demás herramientas informáticas, los humanos están creando y compartiendo recursos digitales (información, expresión creativa, ideas y conocimientos codificados para ser procesados por computadora) que valoran y desean compartir con otros sin restricciones de tiempo ni de espacio. Esta es la prueba que existe un patrimonio digital, es decir, un patrimonio constituido por componentes que poseen muchas características en común y sobre los que pesan muchas amenazas comunes.”[73]

4.4. Desafíos

“Los desafíos son multifacéticos y pueden agruparse en dos categorías:

Vulnerabilidades técnicas

- Medios de almacenamiento, debido al deterioro físico, maltrato, almacenamiento incorrecto y obsolescencia;
- Formatos de archivo y sistemas de compresión, debido a la obsolescencia o demasiada confianza en los formatos de compresión y de archivo patentados y no compatibles;
- Integridad de los archivos, incluyendo la protección del contenido, contexto, fijeza, referencias y procedencia;
- Dispositivos, programas, sistemas operativos, interfaces y protocolos de almacenamiento y procesamiento que cambian a medida que la tecnología evoluciona (con frecuencia con compatibilidad hacia atrás limitada);
- Herramientas de recuperación y procesamiento distribuidas, como por ejemplo textos y aplicaciones Java insertados.

Desafíos administrativos y de organización

- Compromiso institucional de preservación a largo plazo insuficiente;
- Falta de políticas y procedimientos de preservación;
- Escasez de recursos humanos y financieros;
- Intereses variables (y asincrónicos) de quienes tienen participaciones, en la creación, mantenimiento y distribución de colecciones de imágenes digitales;
- Brechas en la memoria institucional debido a la rotación de personal;

- Mantenimiento de registro y metadatos administrativos inadecuados;
- Naturaleza evolutiva de las disposiciones sobre derechos de autor y uso justo que se aplican a las colecciones digitales”[70].

4.5. Estrategias técnicas

Si bien se han propuesto muchas estrategias de preservación digital, ninguna de ellas es apropiada para todos los tipos de datos, situaciones o instituciones. A continuación se describe los puntos básicos a seguir, comenzando por un vocabulario de términos a tener en cuenta .

“Conservación. Es la parte de la gestión de documentos digitales que trata de preservar tanto el contenido como la apariencia de los mismos. Si bien no hay un acuerdo en cuanto a la definición de qué se considera conservación a largo plazo, el lapso de tiempo tiene que presuponerse lo suficientemente amplio como para implicar cambios tanto en la tecnología como en la comunidad de usuarios. Nuevas versiones de bases de datos, hojas de cálculo y procesadores de texto se pueden esperar al menos cada dos o tres años, con correcciones y actualizaciones incluso más a menudo. En general, preservar la apariencia de un documento digital es difícil cuando se trata de texto, pero es casi imposible cuando se trata de entornos multimedia, donde hay una intensa interrelación entre hardware/software y contenidos.

Copias de seguridad. Se refiere al proceso de hacer duplicados exactos del objeto digital. Aunque es un componente esencial de todas las estrategias de preservación, las copias de seguridad en sí mismas no son una técnica de mantenimiento a largo plazo, ya que se ocupa exclusivamente con la cuestión de pérdida de datos debido a un fallo de hardware, bien debido a causas normales, bien a desastres naturales bien

a destrucción malintencionada. En ocasiones, se combina con almacenamiento remoto de tal forma que el original y las copias no estén sujetas a los mismos eventos desastrosos. Las copias de seguridad deberían ser consideradas la estrategia de mantenimiento mínima para incluso los materiales más efímeros y con menos valor que dispongamos.

Actualización. Se refiere a la copia de información digital de un soporte de almacenamiento a largo plazo a otro del mismo tipo, sin ningún cambio en los documentos (por ejemplo, la copia de un viejo CD-RW a otro nuevo).

Metadatos. En el capítulo tres de este trabajo se ha hablado extensamente sobre la necesidad de los metadatos. También en el momento de la conservación existe un consenso entre los expertos al afirmar que es una buena práctica la creación de metadatos en el momento mismo de la producción de los documentos o, como mínimo, crear unos metadatos básicos, que luego serán aumentados en la fase de catalogación e identificación. Además de los metadatos dirigidos a describir el objeto digital y permitir así su recuperación, existen los metadatos dirigidos a la conservación del material. Los metadatos de conservación describen los medios para proporcionar acceso a los datos, junto con aquellos elementos de metadatos requeridos para gestionar los procesos de conservación.

La información que se necesita para compilar unos metadatos destinados a la conservación habitualmente se divide en dos clases en línea con lo establecido en el *Reference Model for an Open Archival Information System* (OAIS)]:

- Información sobre el contenido, consistente en detalles sobre la naturaleza técnica del objeto que indica al sistema cómo

representar los datos con un tipo específico y un formato. A medida que las tecnologías de acceso cambian, estos metadatos de representación también cambian y necesitan ser actualizados.

- Información descriptiva para la conservación, consistente en otra información necesaria para la gestión y uso de los objetos a largo plazo, incluyendo identificadores y detalles bibliográficos, información sobre los propietarios del objeto, gestión de derechos de acceso, historia, contexto incluyendo la relación con otros objetos y la información sobre la validación de los formatos. Obviamente algunos de estos metadatos pueden referirse a otros objetos tales como herramientas de software y especificaciones de formatos que deben a su vez ser conservados. La naturaleza interdependiente de los materiales digitales significa que los programas de conservación a menudo necesitan gestionar redes de objetos digitales enlazados junto con sus metadatos.

Aun no existen estándares aceptados que definan esquemas de metadatos destinados a la conservación, por lo tanto las instituciones interesadas tendrán que elegir entre aceptar (y posiblemente adaptar) alguno de los modelos que están siendo usados por otros o bien diseñar un esquema propio (bien sea como una solución completa de futuro o uno intermedio de carácter mínimo hasta que surja alguna norma que lo sustituya).

A medida que los proveedores de software comiencen a incorporar XML y arquitecturas RDF (*Resource Description Framework*) en sus procesadores de texto y bases de datos, la creación de metadatos como parte integrante del proceso de creación será considerablemente más fácil.

Preservación de la tecnología. Se basa en la preservación del entorno técnico que hace funcionar el sistema, incluyendo sistemas operativos,

software de aplicaciones original, controladores de medios, etc. En cierta forma se trata más de un tipo de estrategia de recuperación después de un desastre de objetos digitales que no han estado sometidos a una conservación adecuada. Ofrece el potencial de tratar con la obsolescencia de los soportes, asumiendo que esos soportes no se han deteriorado más allá de su legibilidad. Puede aumentar el acceso a soportes y formatos de ficheros obsoletos, pero en última instancia es un callejón sin salida, pues ninguna tecnología puede mantenerse funcional de forma indefinida. Esta estrategia no puede ser llevada a cabo por una institución a título individual debido a los altos costes que puede suponer tanto en equipamiento como en personal.

Migración. Se utiliza para copiar o convertir datos desde una tecnología a otra, tanto si se trata de hardware como de software, conservando las características esenciales de los datos. Esta definición captura la esencia y la ambigüedad de la migración. En algunas ocasiones se utiliza como sinónimo de actualización, pero migración representa un concepto mucho más rico y amplio que actualización. Se trata de un conjunto de tareas organizadas destinadas a conseguir la transferencia periódica de materiales digitales desde una generación tecnológica a la siguiente. El propósito de la migración es preservar la integridad de los objetos digitales y mantener la posibilidad por parte de los usuarios de recuperar, visualizar y utilizarlos en una perspectiva de constante cambio tecnológico. La migración incluye la actualización como un medio de conservación digital pero difiere de ella en el sentido de que no siempre es posible hacer una copia digital exacta de un objeto digital cuando el hardware y el software cambian y además deben mantener la compatibilidad del objeto con la nueva generación de tecnología. Si bien las empresas desarrolladoras de software proporcionan estrategias de migración o compatibilidad hacia atrás para algunas generaciones de sus

productos, esto puede no ser verdad más allá de dos o tres generaciones. No obstante, la migración no se garantiza para todos los tipos de datos y se convierte en particularmente poco fiable si el producto de información ha utilizado complicados componentes o características de software.

En estos casos no suele haber compatibilidad hacia atrás y si la hay lo que se produce es una pérdida en la integridad de los contenidos.

Utilización de estándares. Se puede afirmar que la utilización de estándares es al software lo que los soportes perdurables son al hardware. Se trata de buscar una forma de codificar y formatear los objetos digitales adhiriéndose a estándares reconocidos y favoreciéndolos en lugar de los más esotéricos y menos soportados. Presupone que tales estándares perdurarán y que los problemas de compatibilidad que resulten de la evolución del entorno informático (aplicaciones, sistemas operativos) serán solventados debido a la continua necesidad de acomodar la norma dentro de los nuevos entornos. Por ejemplo, si JPEG2000 se convierte en un formato ampliamente utilizado, el elevado número de usuarios garantizará que el software para codificar y visualizar imágenes JPEG2000 se actualizará para responder a las necesidades de nuevos sistemas operativos, etc.

Emulación. La emulación encapsula el comportamiento del software o hardware junto con el objeto digital mismo. Está siendo considerada como una alternativa a la migración. Por ejemplo, un documento en MS Word 2000 podría llevar incorporados metadatos que informaran sobre cómo reconstruir el documento y el propio entorno del software al nivel más bajo de bits y bytes. Una alternativa a la emulación documento a documento, es la creación de un registro que identifique unívocamente entornos de hardware y software y proporcione información sobre cómo recrear dicho entorno para preservar el uso del objeto digital.

En estos momentos no existe ningún sistema que proporcione documentación exhaustiva y la información de emulación requerida para que este modelo sea funcional, particularmente para permitir a un archivo tratar con la variedad de viejas tecnologías.

Almacenamiento. El almacenamiento es a menudo tratado como un estado pasivo en el ciclo de vida, pero los soportes de formatos de almacenamiento van cambiando. La solución más común a este problema de cambiar los medios de almacenamiento es la migración a nuevos sistemas de almacenamiento. Esto es caro y siempre está presente el problema de la pérdida de datos o problemas con la calidad cuando se realiza la transformación. Establecer algoritmos para comprobar la exactitud e integridad de la migración es extremadamente importante.

La **arqueología digital** incluye métodos y procedimientos para rescatar contenidos de medios dañados o de entornos de hardware y software obsoletos o dañados”[69].

4.6. Estrategias organizacionales

“Las soluciones técnicas por si solas no son suficientes para asegurar la larga duración de los recursos digitales. Se recurre a un enfoque holístico, dado que éste reconoce las interdependencias entre componentes técnicos y de organización. Entre los asuntos que se deben tratar en tal estrategia se encuentran las necesidades de contratación de personal y de capacitación, los requisitos financieros, los criterios de reelección, y las necesidades de metadatos de preservación.

Los siguientes programas son ejemplos de enfoques prometedores y prácticos a la preservación digital:

El modelo de referencia [OAIS](#) (*Open Archival Information System* - Sistema de Información de Archivo Abierto) proporciona un marco para la preservación y acceso digital a largo plazo, incluyendo terminología y conceptos para describir y comparar arquitecturas de archivo. Tanto el proyecto NEDLIB como Cedars 1 han adoptado el modelo de referencia OAIS como base para sus exploraciones.

El proyecto [Cedars 1](#) (*CURL Exemplars in Digital Archives* - Modelos CURL en los Archivos Digitales) tiene como objetivo producir marcos estratégicos para las políticas de gestión de colecciones digitales, y promover métodos adecuados para la preservación a largo plazo de diferentes clases de recursos digitales, incluyendo la creación de los metadatos apropiados.

La [Biblioteca Europea de Depósito en Red](#) (*Networked European Deposit Library, NEDLIB*) es un proyecto en colaboración de las bibliotecas nacionales europeas para construir un marco para una nueva biblioteca de depósito en red. Entre los asuntos esenciales que explora se encuentran procedimientos de mantenimiento de archivos y el enlace entre los requisitos de metadatos y las estrategias de preservación.

El proyecto [PANDORA](#) (*Preserving and Accessing Networked Documentary Resources of Australia* - Preservación y Acceso a Recursos Documentales en Red de Australia) ha establecido en forma exitosa un archivo de publicaciones australianas seleccionadas online, ha desarrollado varias políticas y procedimientos de preservación digital, ha redactado un modelo de datos lógico para los metadatos de preservación, y ha esbozado una propuesta para realizar un acercamiento nacional a la preservación a largo plazo de estas publicaciones"[70].

Capítulo V: Revisión de estándares

5.1. Introducción

El estándar de datos sirve para escribir cualquier tipo de información en el mismo formato, se puede escribir en un formato plano permitiendo a las aplicaciones beneficiarse de la interoperabilidad entre sistemas que intercambian información, facilitan el procesamiento automatizado de los recursos web. Por ejemplo: para transacciones comerciales, uso compartido de bases de datos, intercambio de información, etc. Las normas proporcionan un medio para hacer explícito ciertos rasgos de un texto de tal modo que facilite su procesamiento por programas informáticos ejecutados de diversas máquinas, esto se denomina marcado o codificado.

En este capítulo se describen diversos estándares de transferencia de datos, estándares de metadatos y las Normas del Text Encoding Initiative (TEI).

5.2. Estándares para transferencia de datos

5.2.1. Lenguajes de marcado

Un lenguaje de marcado o lenguaje de marcas es una forma de codificar un documento que, junto con el texto, incorpora etiquetas o marcas que contienen información adicional acerca de la estructura del texto o su presentación.

5.2.2. SGML (Standard Generalized Markup Language, ISO 8879)

Es el estándar internacional para la definición de la estructura y el contenido de diferentes tipos de documentos electrónicos. Es un metalenguaje que nos permite definir lenguajes para definir la estructura y el contenido de los documentos.

Un ejemplo de DTD es por ejemplo la que define cómo tendrán que ser los documentos HTML. Por lo tanto, el HTML es un tipo de documento SGML que se utiliza en la Web, y esto es importante, ya que aquí radica su principal diferencia con el XML. El XML no es ningún tipo de documento SGML, sino que es una versión abreviada de SGML optimizada para su utilización en Internet, para mayor comprensión vea la [Figura 2]. Esto significa que con él vamos a poder definir nuestros propios tipos de documentos (podremos definir nuestras propias etiquetas).

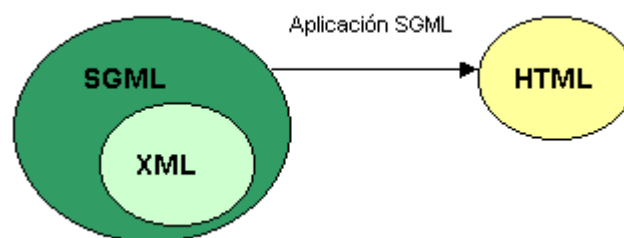


Figura 2. Diferencia entre el SGML, XML y HTML

5.2.3. Lenguaje de Marcado para Texto (HTML - HyperText Markup Language)

Lenguaje de marcas hipertextuales, lenguaje de marcación diseñado para estructurar textos y presentarlos en forma de hipertexto, que es el formato estándar de las páginas web. Fue creado por Berner Lee en base al SGML, Gracias a Internet y a los navegadores del tipo Internet Explorer, Opera, Firefox o Netscape, el HTML se ha convertido en uno de los formatos más populares que existen para la construcción de documentos y también de los más fáciles de aprender . Sin embargo a medida que internet ha ido madurando, las necesidades también han ido evolucionando hacia un lenguaje que pueda usarse para requerimientos más complejos y de gran escala. Por ello vinieron surgiendo problemas del HTML como:

- El contenido se mezcla con los estilos que se le quieren aplicar.
- No permite compartir información con todos los dispositivos, como pueden ser ordenadores o teléfonos móviles.
- La presentación en pantalla depende del visor que se utilice.

5.2.4. Lenguaje de Marca Extensible (XML - Extensible Markup Language)

En el año 1998 World Wide Web (W3C) empezó y continúa, en el desarrollo de XML que fuera mas sencillo de utilizar que el SGML por ello XML es algo así como SGML simplificado, es un formato basado en texto, específicamente diseñado para almacenar y transmitir datos. Un origen XML se compone de elementos XML, cada uno de los cuales consta de una etiqueta de inicio (<title>), de una etiqueta de fin (</title>) y de los datos comprendidos entre ambas etiquetas (el contenido). Un documento XML contiene texto anotado por etiquetas. Admite un conjunto ilimitado

de etiquetas, no para indicar el aspecto que debe tener algo, sino lo que significa. Por ejemplo: un elemento XML puede estar etiquetado como precio, número de pedido o nombre. El autor del documento es quien decide qué tipo de datos va a utilizar y qué etiquetas son las más adecuadas. Los documentos XML son fáciles de crear.

El XML se creó para que cumpliera varios objetivos tales como:

- Que fuera idéntico a la hora de servir, recibir y procesar la información que el HTML, para aprovechar toda la tecnología implantada para este último.
- Que fuera formal y conciso desde el punto de vista de los datos y la manera de guardarlos.
- Que fuera extensible, para que lo puedan utilizar en todos los campos del conocimiento.
- Que fuese fácil de leer y editar.
- Que fuese fácil de implantar, programar y aplicar a los distintos sistemas.

El XML se puede usar para infinidad de trabajos y aporta muchas ventajas en amplios escenarios. Veamos algunas ventajas del XML en algunos campos prácticos.

- **Comunicación de datos.** Si la información se transfiere en XML, cualquier aplicación podría escribir un documento de texto plano con los datos que estaba manejando en formato XML y otra aplicación recibir esta información y trabajar con ella.
- **Migración de datos.** Si tenemos que mover los datos de una base de datos a otra sería muy sencillo si las dos trabajasen en formato XML.

- **Aplicaciones web.** Hasta ahora cada navegador interpreta la información a su manera y los programadores del web tenemos que hacer unas cosas u otras en función del navegador del usuario. Con XML tenemos una sola aplicación que maneja los datos y para cada navegador o soporte podremos tener una hoja de estilo o similar para aplicarle el estilo adecuado. Si mañana nuestra aplicación debe correr en WAP solo tenemos que crear una nueva hoja de estilo o similar.
- **Es internacional,** ya que se basa en el Unicode (estándar que proporciona un número único para cada carácter, sin importar la plataforma, programa o idioma).
- Permite separar contenido y presentación.
- Puede estar estructurado utilizando DTD o XML Schema.
- Puede ser contenedor de datos

5.2.4.1 Estructura de un documento XML

Un documento XML tiene dos estructuras, una **lógica** y otra **física**. Físicamente, el documento está compuesto por unidades llamadas entidades. Una entidad puede hacer referencia a otra entidad, causando que esta se incluya en el documento. Cada documento comienza con una entidad documento, también llamada raíz. Lógicamente, el documento está compuesto de declaraciones, elementos, comentarios, referencias a caracteres e instrucciones de procesamiento, todos los cuales están indicados por una marca explícita. Las estructuras lógica y física deben encajar de manera adecuada.

Los documentos XML se dividen en dos grupos, **documentos bien formados** y **documentos válidos**.

- **Bien formados:** Son todos los que cumplen las especificaciones del lenguaje respecto a las reglas sintácticas sin estar sujetos a unos elementos fijados en un DTD.
- **Válidos:** Además de estar bien formados, siguen una estructura y una semántica determinada por un DTD: sus elementos y sobre todo la estructura jerárquica que define el DTD, además de los atributos, deben ajustarse a lo que el DTD dicte.

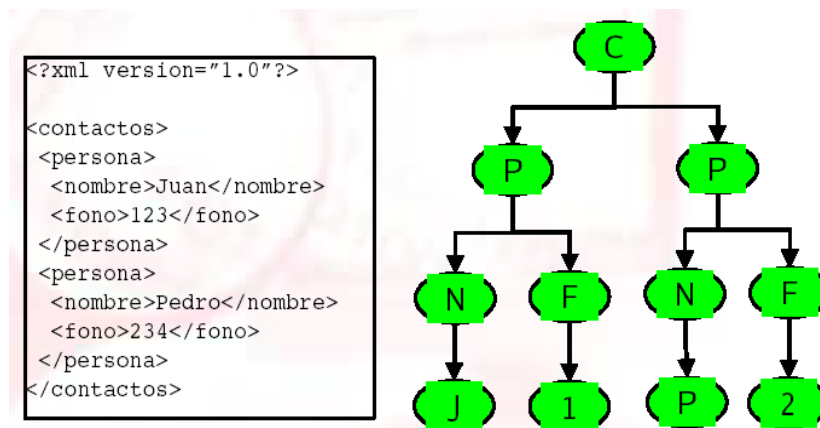


Figura 3: Estructura del XML

Los documentos XML deben seguir una estructura estrictamente jerárquica con lo que respecta a las etiquetas que delimitan sus elementos tal como se observa en la [Figura 3]. En XML una etiqueta debe estar correctamente "incluida" en otra. Además, los elementos con contenido, deben estar correctamente "cerrados". En el siguiente ejemplo, la primera línea sería incorrecta en XML, no así la segunda:

```

<LI>HTML <B>permite <I>esto</B></I>.
<LI>En XML la <B>estructura <I>es</I> jerárquica</B></LI>
  
```

5.2.4.2. Diferencias entre XML Y HTML

Al igual que el HTML, XML delimita su contenido con etiquetas sin embargo existen diferencias significativas, que pasamos a detallar a continuación :

- El HTML se preocupa por formatear datos y para ello son las etiquetas que tiene el lenguaje, para formatear la información que se desea mostrar.
- El XML se preocupa por estructurar la información que pretende almacenar. La estructura marca la lógica propia de la información.
- El desarrollo del HTML estuvo marcado la competencia entre los distintos visores del mercado. Cada uno quería ser el mejor e inventaba etiquetas nuevas que a la larga entraban a formar parte del estándar del W3C, como la etiqueta <FRAME>.
- El desarrollo del XML está siendo llevado a cabo con mayor intensidad siempre ajustado a lo que marca el estándar que desarrolla el W3C, entidad que está desarrollando el XML con más diligencia que las empresas con intereses particulares.
- Procesar la información en HTML es inviable, por estar mezclada con los estilos y las etiquetas que formatean la información.
- En XML se puede procesar la información con mucha facilidad, porque todo está ordenado de una manera lógica, así mismo el formateo de la información para que se pueda entender bien por el usuario es viable a través de un pequeño procesamiento, a través de hojas de estilos o similares.

A manera de resumen se observa en la [Tabla 1], algunas de las diferencias con respecto a los lenguajes de marcas que se han mencionado:

	HTML/DHTML	XML	SGML
Gramática	Fija y no ampliable	Extensible	Extensible
Estructura	Monolítica	Jerárquica	Jerárquica
N° de marcas	Fijas	Sin límite	Sin límite
Complejidad	Baja	Mediana	Alta
Diseño de páginas	Fijado por tags. Etiquetas con atributos CSS en DHTML	CSS o XSL	DSSSL
Enlaces	Simples enlaces	Poderosos enlaces (XLL)	HyTime
Exportabilidad (formatos/aplicaciones)	No	Sí	Sí
Validación	Sin validación	Pueden validarse	Obligatorio DTD
Búsquedas	Simple y a veces resuelta por scripts o CGI	Potente búsqueda. Con capacidad para personalizarla	Son posibles potentes búsquedas.
Indización/Catalogación de páginas web	Sólo lo permite los atributos de la etiqueta <META>, e implementaciones como DC.	Una descripción abierta y personalizable con el RDF.	Algún proyecto como TEI, DLI, etc.

Tabla 1: Diferencias entre HTML, XML y SGML

En la [Figura 4], se observa el esquema de las relaciones entre los metalenguajes:

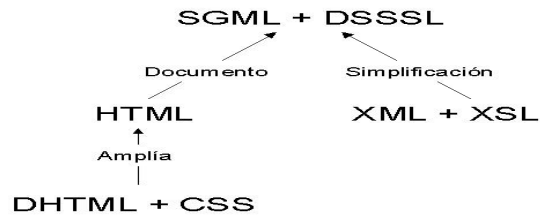


Figura 4: Esquema de relaciones entre lenguaje

5.2.4.3. Declaración de Tipo de Documento (DTD) y Esquema (Eschema)

Son unos lenguajes propios que sirven para definir otros lenguajes, es decir, son metalenguajes. Se definen especificando qué etiquetas podemos o debemos encontrarnos en los documentos XML, en qué orden, dentro de qué otras, además de especificar los atributos que pueden o deben tener cada una de las etiquetas. Hay dos metalenguajes con los que definir los lenguajes que podemos obtener a partir de XML, el DTD y el XML Schema.

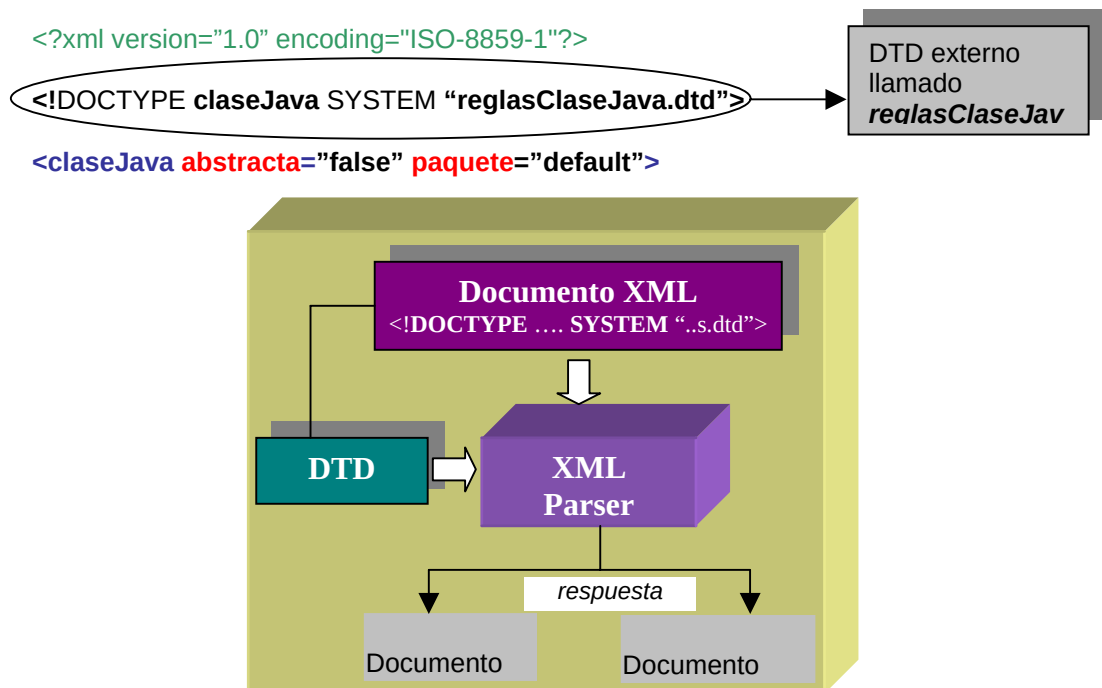


Figura 5: Esquema de validación

Como se vé en la [Figura 5], cuando un Parser procesa un documento XML verifica si tiene asociado un DTD. En caso afirmativo verifica si el documento satisface las reglas que se han definido en el DTD o Schema y sigue procesando. De encontrar un error el proceso de parsing termina. Un Parser no hace nada para corregir un error. La especificación del procesamiento de documentos XML dada por la W3C deja claro que al encontrar un error el proceso de parsing finaliza. Si una instancia satisface las reglas gramaticales pasa a ser válido, caso contrario el Parser informa del error ocurrido.

DTD y XML Schema nos permiten comprobar la integridad de los datos en cualquier momento. Nos sirven para tomar un documento XML y comprobar que los datos que él incluye son válidos, comprobando si lo que tenemos en el XML concuerda con lo que tendríamos que tener.

5.3. Estándares de metadatos

5.3.1. Estándares World Wide Web Consortium(W3C)

El World Wide Web Consortium (<http://www.w3c.org/>) ha mostrado un fuerte interes en los metadatos al desarrrrollar RDF (Resource Description Framework) y las especificaciones para la plataforma para la selección de contenidos en internet (PICS).

RFD es una base para procesar metadatos; proporciona interoperabilidad entre aplicaciones que intercambian información legible por máquina en la Web , utiliza XML para intercambiar descripciones de recursos, es un modelo de datos y sintaxis para los metadatos de los recursos web [Candam et al. 2001]. Un recurso puede ser una pagina web, un sitio web completo o cualquier elemento en la web que tenga información en cualquier formato. RDF permite la codificación, el intercambio y la

reutilización de metadatos estructurados . Permite la interoperabilidad de los metadatos por medio del diseño de mecanismos que soporten convenios comunes de semántica , sintaxis y estructura.

PICS fue desarrollado por W3C para permitir la asociación de etiquetas (metadatos) con contenido en internet. Inicialmente fue diseñado para el control del acceso al internet de los niños a ciertos sitios en Internet. Es una Recomendación del W3C para intercambiar descripciones de contenido de las páginas Web y otros materiales. PICS es un predecesor de RDF y es un requisito explícito de RDF que sea capaz de expresar cualquier cosa que se pueda expresar a través de una etiqueta PICS.

5.3.2. Estándar de metadatos Dublin Core (DC)

“El estándar de metadatos Dublin Core es simple pero eficaz conjunto de elementos para describir una amplia gama de recursos de la red. La norma del Dublin Core conlleva dos niveles: Simple y Cualificado.

El Dublin Core Simple contiene quince elementos; divididos en tres grupos:

Contenido: Título, Tema, descripción, fuente, idioma, relación, cobertura.

Propiedad intelectual: autor, editor, colaborador, derechos.

Instanciación: fecha, tipo, formato, identificador

El Dublin Core Cualificado conlleva un elemento adicional, la audiencia (Audience), así como un grupo de elementos de matización (denominados por ello, cualificadores) que refinan la semántica de los elementos de tal forma que pueden ser útiles para la recuperación/localización de recursos en Internet.

La semántica del Dublin core se ha establecido por un grupo internacional e interdisciplinario de profesionales de la biblioteconomía, la Informática, la codificación textual, la comunidad museística, y otros campos teórico-prácticos relacionados.

No existen restricciones fundamentales para los tipos de recursos a los que se les pueden asignar metadatos Dublin Core. Cada elemento es opcional y puede repetirse, los elementos son los siguientes:

1.- Nombre del elemento: Title (Título)

Etiqueta: Title

Definición: Nombre dado al recurso

Comentario: Normalmente, el título será el nombre por el que se conoce formalmente el recurso.

2.- Nombre del elemento: Creator (Creador)

Etiqueta: Creador

Definición: Entidad principal responsable de crear el contenido del recurso.

Comentario: Ejemplo de Creador de un recurso pueden ser, una persona, una organización, o un servicio. Normalmente el nombre de un creador debe usarse para indicar la entidad.

3.- Nombre del elemento: Subject (Materia)

Etiqueta: Subject y Keywords

Definición: Tema del contenido del recurso.

Comentario: Normalmente, la materia se expresará como palabras clave, frases clave o códigos de clasificación que describan un tema/asunto del recurso. La mejor práctica recomendada es

seleccionar un valor de un vocabulario controlado o de un esquema de clasificación formal.

4.- Nombre del elemento: Description (Descripción)

Etiqueta: Description

Definición: Explicación del contenido del recurso.

Comentario: Ejemplos de descripción son, aunque no están limitados a: un resumen, tabla de contenidos, referencia a una representación gráfica del contenido o una explicación en texto libre sobre el contenido.

5.-Nombre del elemento: Publisher (Editor)

Etiqueta: Publisher

Definición: Entidad responsable de que el recursos esté disponible

Comentario: Ejemplos de editor son: una persona, una organización, o un servicio. Normalmente, el nombre de un editor debe utilizarse para indicar la entidad.

6.- Nombre del elemento: Contributor (Colaborador)

Etiqueta: Contributor

Definición: Entidad responsable de realizar contribuciones al contenido de un recurso.

Comentario: Ejemplos de colaborador pueden ser: una persona, una organización o un servicio. Normalmente el nombre de un colaborador debe utilizarse para indicar una entidad.

7.- Nombre del elemento: Date (Fecha)

Etiqueta: Date

Definición: Fecha de una circunstancia relativa al ciclo de vida de un recurso.

Comentario: Normalmente, la fecha se asociará con la creación o la disponibilidad de un recurso. La mejor práctica recomendada para codificar el valor de la fecha se define en el perfil *ISO 8601*, e incluye fechas en la forma AAAA-MM-DD.

8.-Nombre del elemento: Type (Tipo de recurso)

Etiqueta: Resource Type

Definición: Naturaleza o género del contenido del recurso.

Comentario: El tipo se refiere a términos que describen categorías generales, funciones, géneros o niveles de agregación del contenido. La práctica recomendada en este sentido, es seleccionar un valor de un vocabulario controlado (por ejemplo, del Vocabulario de tipos de la DCMI (Dublín Core Metadata Initiative)).

9.-Nombre del elemento: Format (Formato)

Etiqueta: Format

Definición: Manifestación física o digital de un recurso

Comentario: Normalmente, el formato puede incluir tipos de medios o dimensiones de un recurso. El formato puede usarse para identificar el software, hardware, u otros equipamientos necesarios para visualizar u operar el recurso. Ejemplos de dimensiones pueden ser el tamaño o la duración. La práctica recomendada en este caso es seleccionar el valor de un vocabulario controlado (por ejemplo, la lista de Tipos de Medios en Internet (MIME) que define los formatos de medios informáticos).

10.-Nombre del elemento: Identifier (Identificador)

Etiqueta: Resource Identifier

Definición: Referencia inequívoca a un recurso dentro de un contexto dado.

Comentario: La mejor práctica recomendada es identificar el recurso por medio de un *string* [serie de caracteres manipulados como un grupo] o número adaptado a un sistema formal de identificación. Algunos formatos de identificación formal de recursos son, entre otros, el Identificador Uniforme de Recursos (URI) (que incluye el Localizador Uniforme de Recursos (URL)), el Identificador de Objetos Digitales (DOI) y el Número Internacional Normalizado de Libros (ISBN).

11.-Nombre del elemento: Source (Fuente)

Etiqueta: Source

Definición: Referencia a un recurso del cual deriva el recurso actual (que se está describiendo).

Comentario: El recurso actual puede derivar de un recurso Fuente en todo o en parte. La práctica mejor recomendada en este caso es identificar el recurso referenciado por medio de un *string* o número conforme con un sistema de identificación formal.

12.- Nombre del elemento: Language (Idioma)

Etiqueta: Language

Definición: Idioma del contenido intelectual de un recurso.

Comentario: La mejor práctica recomendada es usar la RFC 3066 que, en conjunción con la norma ISO639 define dos y tres etiquetas de idioma principal con subetiquetas opcionales. Por ejemplo, se incluye "en" o "eng" para el inglés, "akk" para el acadio, y "en-GB" para el inglés usado en el Reino Unido.

13.-Nombre del elemento: Relation (Relación)

Etiqueta: Relation

Definición: Referencia a un recurso relacionado.

Comentario: La práctica recomendada es identificar los recursos referenciados mediante un *string* [conjunto de caracteres manipulados como un grupo] o un número conforme a un sistema de identificación formal.

14.-Nombre del elemento: Coverage (Cobertura)

Etiqueta: Coverage

Definición: La magnitud o el alcance del contenido de un recurso.

Comentario: Normalmente, la cobertura incluirá la localización espacial (un nombre de un lugar o unas coordenadas geográficas), periodo temporal (una expresión que identifica un período, fecha o rango de fecha) o jurisdicción (por ejemplo una denominación de una entidad administrativa). La práctica recomendada es seleccionar un valor de un vocabulario controlado (por ejemplo, del Tesoro de Nombres Geográficos [TGN]) y usar, cuando sea oportuno, nombres de periodos de tiempo o de lugares, mejor que identificadores numéricos, como un conjunto de coordenadas o rangos de fecha.

15.- Nombre del elemento: Rights (Derechos)

Etiqueta: Rights Management

Definición: Información sobre los derechos legales que afectan al uso del recurso.

Comentario: Normalmente, los derechos contendrán una declaración de gestión de derechos para el recurso, o referenciarán un servicio que proporcione dicha información. La información sobre los derechos normalmente abarca los derechos de Propiedad

Intelectual (IPR), Copyright, y varios derechos relacionados con la propiedad. Si no consta el elementos de Derechos, no se deben hacer asunciones sobre ningún derecho contenido en el recurso o entorno a él. " [40]

5.4. Iniciativa de texto codificado (text encoding initiative TEI)

5.4.1. Introducción

Las Normas del Text Encoding Initiative (TEI) se usan para intercambiar información almacenada en un formato electrónico. En ellas se enfatiza el intercambio de información textual, pero también se hace referencia a otras formas de información como son las imágenes y el sonido. Las Normas son del mismo modo aplicables a la creación de nuevos recursos y al intercambio de los ya existentes.

El TEI apareció, en parte, debido a la gran variedad de esquemas de codificación mutuamente incomprensibles que abundan actualmente en los ámbitos de estudio, y en parte, debido a la creciente variedad de usos que están siendo identificados para los textos en formato electrónico.

Las Normas TEI usan el Standard Generalized Markup Language (SGML) para definir su esquema de codificación. El uso que hace el TEI del SGML es ambicioso en su complejidad y generalidad, pero fundamentalmente no es diferente del de cualquier otro esquema de marcado en SGML. Por lo tanto, cualquier programa preparado para SGML puede procesar los textos que cumplan el TEI.

El TEI está patrocinado por las siguientes organizaciones: Association for Computers and the Humanities, la Association for Computational Linguistics, y la Association for Literary and Linguistic Computing. La

financiación ha sido facilitada, en parte, por el U.S. National Endowment for the Humanities, Directorate General XIII of the Commission of the European Communities, la Fundación Andrew W. Mellon, y el Social Science and Humanities Research Council of Canada. Sus Normas fueron publicadas en mayo de 1994, tras seis años de desarrollo donde participaron cientos de estudiosos de diferentes disciplinas académicas de todo el mundo.

Al comienzo los objetivos generales del TEI fueron definidos por la declaración de clausura de una conferencia de organización que tuvo lugar en Vassar College, N.Y., en noviembre de 1987; estos principios 'Poughkeepsie' fueron más tarde elaborados en series de documentos de diseño. Las Normas, según estos documentos, deben:

- Ser suficientes para representar los rasgos textuales necesarios para la investigación;
- Ser sencillas, claras y concretas;
- Ser fáciles para los investigadores para que las usen sin ningún programa específico;
- Permitir la definición rigurosa y el procesamiento eficaz de los textos;
- Prevenir las extensiones definidas por los usuarios;
- Ajustarse a los estándares existentes y en desarrollo.

El mundo de estudio es amplio y diverso. Para que las Normas sean aceptadas era importante asegurar que:

- El núcleo común de los rasgos textuales fuera fácilmente compartido;
- Otros rasgos específicos fueran añadidos (o eliminados) fácilmente a un texto;

- Fueran posibles múltiples codificaciones semejantes de un mismo rasgo;
- La riqueza del marcado debía ser definida por el usuario, con un mínimo nivel de exigencia;
- Una adecuada documentación del texto y de su codificación debía ser suministrada.

Algunos elementos de la selección TEI es el llamado TEI Lite. Al seleccionar de los varios cientos de elementos SGML definidos por el TEI completo, se ha tratado de identificar un “conjunto inicial” que recoja los elementos que casi cualquier usuario debería conocer. La experiencia al trabajar con el TEI Lite será de gran valor para entender la DTD del TEI completo y para conocer qué partes opcionales de la DTD completa son necesarias para trabajar con tipos de textos específicos.

EL objetivo de “TEI Lite” al definir este subconjunto se puede resumir del siguiente modo:

- Debe incluir la mayoría de las etiquetas principales del TEI, ya que estas contienen elementos importantes para todos los tipos de textos y todos los tipos de trabajo de procesamiento de los textos.
- Debe poder manejar correctamente una variedad de textos bastante amplia, a un nivel de detalle que se dé en la práctica real
- Debe ser útil para la producción de nuevos documentos así como para etiquetar los ya existentes.
- Debe ser utilizable por la mayoría de los programas de SGML ya existentes.
- Debe poder ser obtenido de la completa TEI DTD empleando los mecanismos de extensión descritos en las Normas del TEI.

- Debe ser lo más reducido y sencillo posible siempre y cuando sea coherente con los otros objetivos.

Todos los elementos descritos están totalmente documentados en las normas del TEI, que deben ser consultadas para mayor información sobre los aspectos tratados, y no tratados, aquí.

5.4.2. Estructura de un texto TEI

Todos los textos que cumplan el TEI contienen:

- (a) Un encabezado *TEI* (marcado con el elemento <teiHeader>) y
- (b) la transcripción del propio texto (marcado con el elemento <text>).

El encabezado TEI provee información similar a la de la portada de un texto impreso. Tiene hasta cuatro partes: una descripción bibliográfica del texto electrónico, una descripción de cómo ha sido etiquetado, una descripción no bibliográfica del texto (un perfil del texto), y una revisión de su historia (su creación). Un texto TEI puede ser individual (una única obra) o compuesto (una colección de obras, como por ejemplo una antología). En cualquier caso, el texto puede tener un front o back opcional. En medio está el body, cuerpo de la obra, que, en el caso de un texto compuesto, puede estar formado por groups, cada uno conteniendo a su vez más grupos o textos. Un texto individual se etiquetará siguiendo una estructura genérica como esta:

```
<TEI.2>
  <teiHeader> [ Información del encabezado TEI ] </teiHeader>
  <text>
    <front> [ materia del front ... ] </front>
    <body> [ cuerpo del texto ... ] </body>
    <back> [ materia back ... ] </back>
  </text>
</TEI.2>
```

Un texto compuesto también tiene una materia previa (front) y final (back) opcional. En medio aparecen uno o más grupos de textos, cada uno con sus propios front y back opcionales. Un texto compuesto se etiquetará usando una estructura genérica como esta:

```
<TEI.2>
  <teiHeader> [ información del encabezado del conjunto de
textos] </teiHeader>
<text>
<front> [ materia del front del conjunto de textos ] </front>
<group>
  <text>
    <front> [ materia del front del primer texto ] </front>
    <body> [ materia del body del primer texto ] </body>
    <back> [ materia del back del primer texto ] </back>
  </text>
  <text>
    <front> [ materia del front del segundo texto ] </front>
    <body> [ materia del body del segundo texto ] </body>
    <back> [ materia del back del segundo texto ] </back>
  </text> [ aquí van los demás textos o grupos de textos ]
</group>
<back> [ materia del back del conjunto de textos ] </back>
</text>
</TEI.2>
```

También es posible definir un conjunto de textos TEI, cada uno con su propio encabezado. Este tipo de colección se llama corpus TEI, y toda esta puede tener un encabezado:

```
<teiCorpus>
  <teiHeader>    [ información del encabezado del corpus ]
</teiHeader>
  <TEI.2>
    <teiHeader> [ información del encabezado del primer texto ]
  </teiHeader>
    <text> [primer texto del corpus] </text>
  </TEI.2>
  <TEI.2>
    <teiHeader> [ información del encabezado del primer texto ]
  </teiHeader>
    <text> [second texto del corpus ] </text>
  </TEI.2>
</teiCorpus>
```

Sin embargo, no es posible crear un conjunto de corpora, es decir, serie de elementos `<teiCorpus>` combinados juntos y tratados como un único objeto. Esta es una restricción de la actual versión de las Normas TEI.

La descripción de todos los elementos del TEI se encuentran en la página principal que se encuentra en la bibliografía [62].

Capítulo VI: Recuperación de Información

6.1. Introducción

Las bibliotecas digitales además de contenidos deben ofrecer servicios que guíen y faciliten a los usuarios en la explotación de los mismos. Los sistemas de recuperación de datos tienen como objetivo localizar información en grandes volúmenes de colecciones documentarias que están en formato electrónico. Se les formulan consultas (búsquedas) que expresan que contenidos se desea localizar. Para poder realizar dicho trabajo es necesario que el sistema con anterioridad procese la colección de documentos con el fin de construir estructuras de acceso (índices) que les permitan una localización rápida. Los buscadores del Internet son los sistemas de recuperación más populares pero existen muchos que se pueden usar dependiendo el tipo de requerimiento.

Actualmente aprovechando el formato digital de los libros en las bibliotecas digitales se están implementando los servicios de búsqueda por contenido. El motor de búsqueda “Lucene.net” es uno de los sistemas de recuperación de información (SRI) por contenido por ello la implementaremos en el proyecto.

En este capítulo se describe brevemente la teoría previa para poder entender el funcionamiento de “Lucene.net” y la descripción de la forma de implementación del proyecto.

6.2. Almacenamiento y recuperación de información en datos no estructurados

Los datos no estructurados por lo general son algunos tipos de documentos. Si deseamos realizar una búsqueda, se tendría que revisar si los términos que conforman la búsqueda se encuentran en el documento, esta tarea si es fácil describir pero se puede volver compleja, por ello surgió la necesidad de indexar los documentos.

6.2.1. Indexación de un documento

Indexar un documento es extraer todos los términos contenidos en ella y guardarlo en una lista de tal manera que nos permita ir desde la consulta hasta el documento que lo contiene. Para la indexación, la mayoría de las técnicas de análisis y representación de documentos utilizadas en la actualidad en los sistemas de clasificación de documentos, se basan en criterios fundamentalmente estadísticos, centrados en frecuencias de aparición de términos en documentos.

Dentro de los sistemas de clasificación de documentos se puede distinguir los *sistemas de recuperación de texto*, que seleccionan aquellos textos o documentos que son adecuados a una necesidad del usuario entre un conjunto más amplio, y *sistemas de agrupamiento de textos* que, a partir de un conjunto de textos, construyen subconjuntos de éstos con contenidos semejantes.

Si trabajamos con sistemas de clasificación de documentos, debemos centrarnos en tres puntos fundamentales:

- **Método de representación:** forma en que serán representados internamente los documentos.

- **Método de análisis:** proceso que permitirá obtener las representaciones concretas de los documentos a partir del análisis de su contenido.
- **Método de cálculo de similitud:** entre las representaciones de los documentos y la representación del otro elemento de la operación de clasificación.

En la [Figura 6], se representa la operación de recuperación de textos centrada en los tres puntos descritos. En ella se puede observar que, en este caso, el elemento específico de la operación de clasificación es una consulta de usuario; es decir, que los documentos que se recuperarían serían aquellos cuya representación interna presentase una mayor similitud con la de la consulta del usuario.

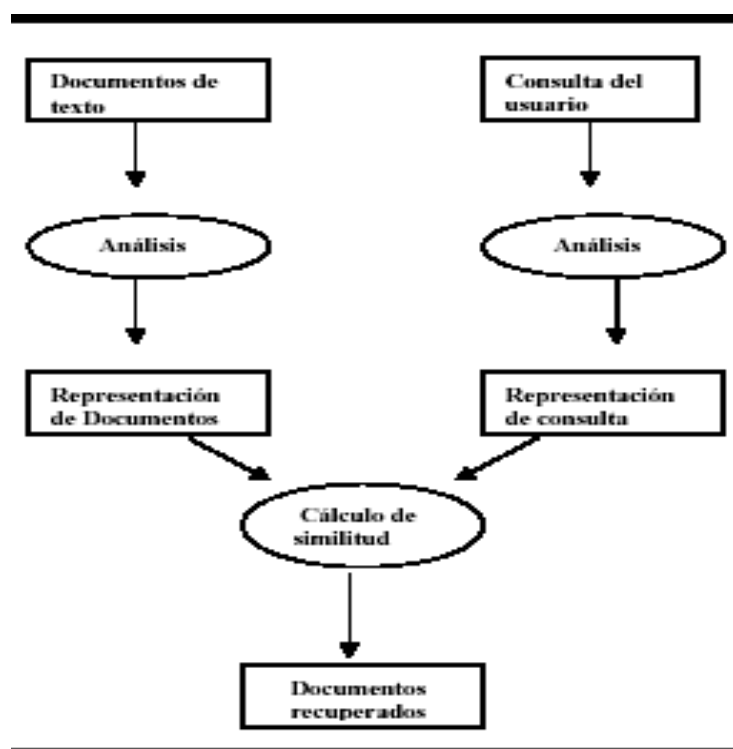


Figura 6: Operación de recuperación de textos.[54]

Una vez introducida la recuperación de textos, diremos que el término de *indexación* hace referencia a un método que engloba la definición de uno de representación y uno de análisis. El término *proceso de indexación* hace

referencia al proceso de análisis de documentos para la obtención de una representación concreta de los mismos.” [54]

“En el *modelo del espacio vectorial* se propone la representación de cada documento mediante un vector cuyos componentes son los pesos asociados a los términos utilizados en la representación. [\[Ver Anexo2\]](#).

Para realizar el proceso de indexación utilizando una representación basada en el espacio vectorial se pueden utilizar los siguientes elementos, que permiten obtener la representación interna de los documentos mediante un análisis automático de su contenido:

1. **Peso de los términos**

El concepto de poder de resolución de un término proporciona una base para los métodos de indexación basados en frecuencia de aparición de términos. El poder de resolución de un término proporciona información acerca de su adecuación como término de indexación.

2. **Listas de parada**

Las listas de parada (*stoplists*) se utilizan en el análisis de los documentos para la eliminación de una serie de palabras que no resultan útiles para la obtención de términos de indexación, por ejemplo, *de, en, el*, etc.

3. **Extracción de raíces**

Los algoritmos de extracción de raíces (*stemming*), o de eliminación de sufijos, se encuentran orientados a obtener un único término a partir de diferentes palabras que constituyen, esencialmente, variaciones morfológicas con un mismo significado.

El resultado del algoritmo debe ser una misma forma regular para las diferentes variantes morfológicas de una palabra, que no tiene por qué ser, necesariamente, la raíz lingüística.

4. Frases de términos

Las frases de términos se orientan a la obtención de términos de indexación con un significado más preciso que el de los términos obtenidos directamente a partir de las palabras individuales. Una frase de términos es una tupla de términos y constituye en sí misma un nuevo término de indexación.

5. Thesaurus

Un thesaurus proporciona una agrupación o clasificación de términos en un determinado dominio o área en categorías denominadas clases. Permite recuperar documentos que son relevantes a la consulta de un usuario, aunque no aparezcan en ellos los términos de la consulta, pero si sinónimos de estos. “ [\[54\]](#)

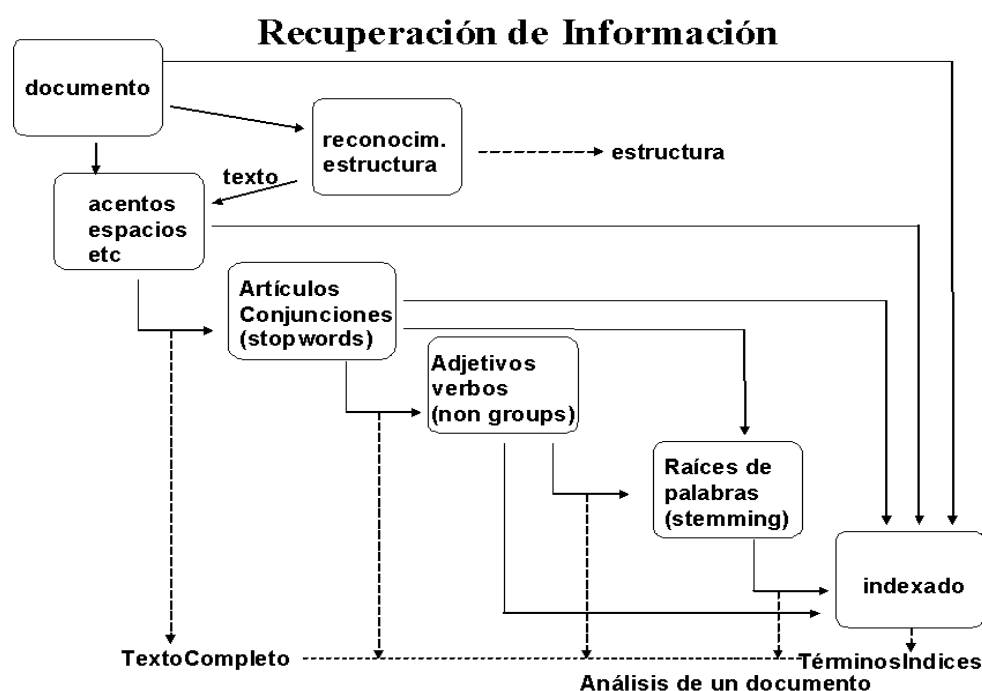


Figura 7: Procesamiento de la información en un SRI.[56]

Para no tener que almacenar e indexar todos los términos de un documento, en la [Figura 7] se muestra algunas técnicas para eliminar aquellos que son innecesarios:

“Lematización: Eliminación de palabras comunes en el lenguaje como artículos, conjunciones, etc. Actualmente muy pocos sistemas lo aplican.

Truncamiento: Dejar las palabras en sus raíces sintácticas, por ejemplo la palabra computación, computacionalmente, computadora son reducidas a “comp” y se pueden hacer búsqueda por cualquiera de los términos indexados.”[54]

6.2.2. Modelo de recuperación de información

“Cuando un usuario realiza una búsqueda se encuentra con gran cantidad de documentos que satisfacen su consulta, aquí surge la

“relevancia”, permitiendo ordenar los resultados considerando la utilidad o similitud que tiene cada documento con la consulta del usuario.

Para determinar la relevancia, con el tiempo han ido optimizándose y surgiendo nuevos modelos, hoy en día existen diversos tipos de modelos de recuperación de datos que se centran en características de los términos de los documentos, sus frecuencias dentro de cada documento y dentro de la colección.”[56]

En esta sección tal como se observa en la [Figura 8], a manera de información en el gráfico se detalla un resumen de los modelos de recuperación de información para mayor detalle vea el [Anexo 2].

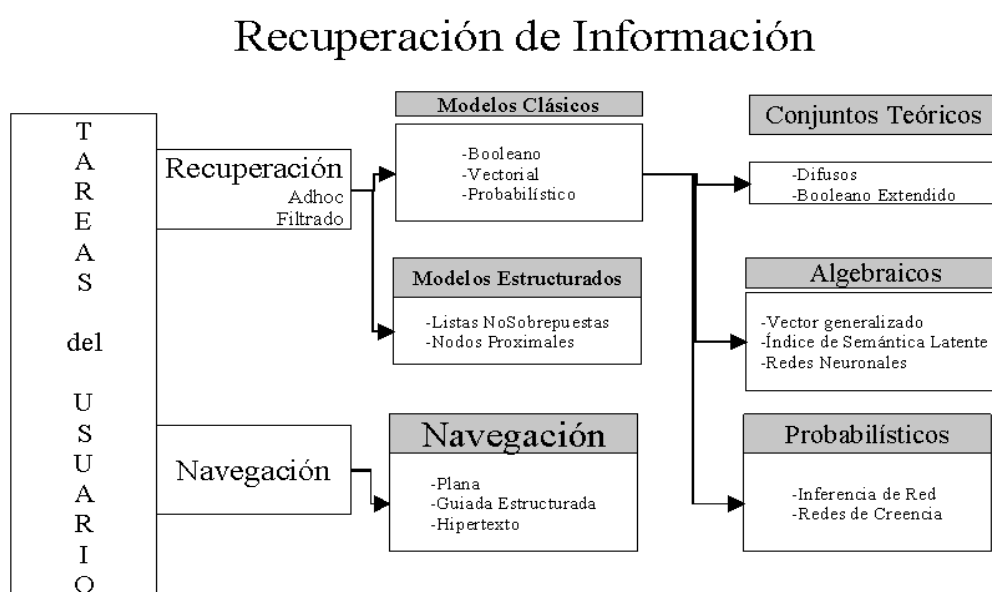


Figura 8: Modelos de recuperación de información. [56]

“Actualmente el modelo de espacios vectoriales, es el modelo más popular (con variantes) ya que permite discriminar correctamente entre documentos. Fue Propuesto en 1975, también llamado solamente “modelo vectorial”, tiene las siguientes características:

- Generaliza el modelo booleano al manejar pesos no binarios para los términos de cada documento y de la consulta (query)
- Determina la similitud entre la consulta y cada documento calculando el ángulo entre sus vectores
- La idea principal es que dos documentos son similares si sus vectores apuntan hacia la misma dirección
- Un aspecto importante es la determinación del peso de cada término, típicamente basado en su frecuencia:
- El peso aumenta entre más ocurre en un documento dado.
- El peso disminuye entre más aparece en los demás documentos.”[56]

En resumen, la arquitectura completa de un sistema de recuperación de información [Figura 9] se basa en el indexamiento de términos y el ordenamiento por relevancia de los resultados en base a los distintos modelos, existiendo una retroalimentación del usuario que permite refinar las consultas y obtener documentos más próximos a las necesidades de cada persona.

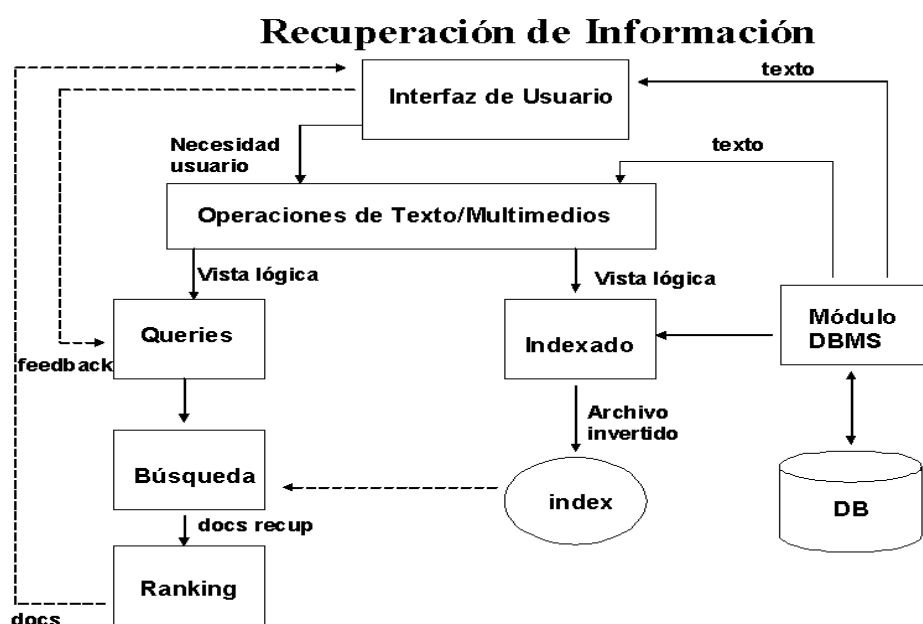


Figura 9: Arquitectura de un SRI. [56]

Existen muchas implementaciones de los modelos de recuperación, entre las más populares se presentan en la [Tabla 2].

Producto	Lenguaje	Modelos	Indexamiento	API desarrollo	Licencia
Managing Gigabytes	C/C++, envoltura Java	Vectorial	Propietario	Ninguno	Open Source
Xapian	C/C++	Probabilístico	Propietario	Limitado	Open Source
Lucene	Java	Vectorial	Propietario	Abierto y extendible	Open Source
MySQL FullText	C/C++ Java (SMART)	Vectorial	Propietario ISAM	SQL	Open Source
Egothor	Java	Booleano Extendido	Propietario	Ninguno	Open Source

Tabla 2: Comparación de herramientas de recuperación de información.
[56]

Para el desarrollo del protipo he optado por usar el SRI Lucene.net, la cual está implementada para la plataforma de .NET con el lenguaje de programación C#, su desarrollo está basada en el modelo vectorial. Lucene.net es la herramienta más adecuada para el tipo de datos que se quiere recuperar e indexar (XML, HTML, PDF) en este trabajo de tesis. Encambio para almacenar y recuperar los metadatos se ha usado el manejador de datos MSSQL 2000 server.

6.3. Lucene

“Lucene es una herramienta que permite tanto la indexación cómo la búsqueda en el contenido de los documentos, es de código abierto, originalmente implementada en Java por Doug Cutting. Está apoyado por el Apache Software Foundation y se distribuye bajo la Apache Software License. Creada bajo una metodología orientada a objetos e implementada completamente en Java, no se trata de una aplicación que pueda ser descargada, instalada y ejecutada sino de una API a través de la cual se pueden añadir capacidades de indexación y búsqueda a cualquier sistema.” [56]

“El centro de la arquitectura lógica de Lucene se encuentra el concepto de Documento (Document) que contiene Campos (Fields) de texto. Esta flexibilidad permite a Lucene ser independiente del formato del fichero. Textos que se encuentran en PDFs, páginas HTML, documentos de Microsoft Word, así como muchos otros pueden ser indexados mientras que se pueda extraer información de ellos.” [56]

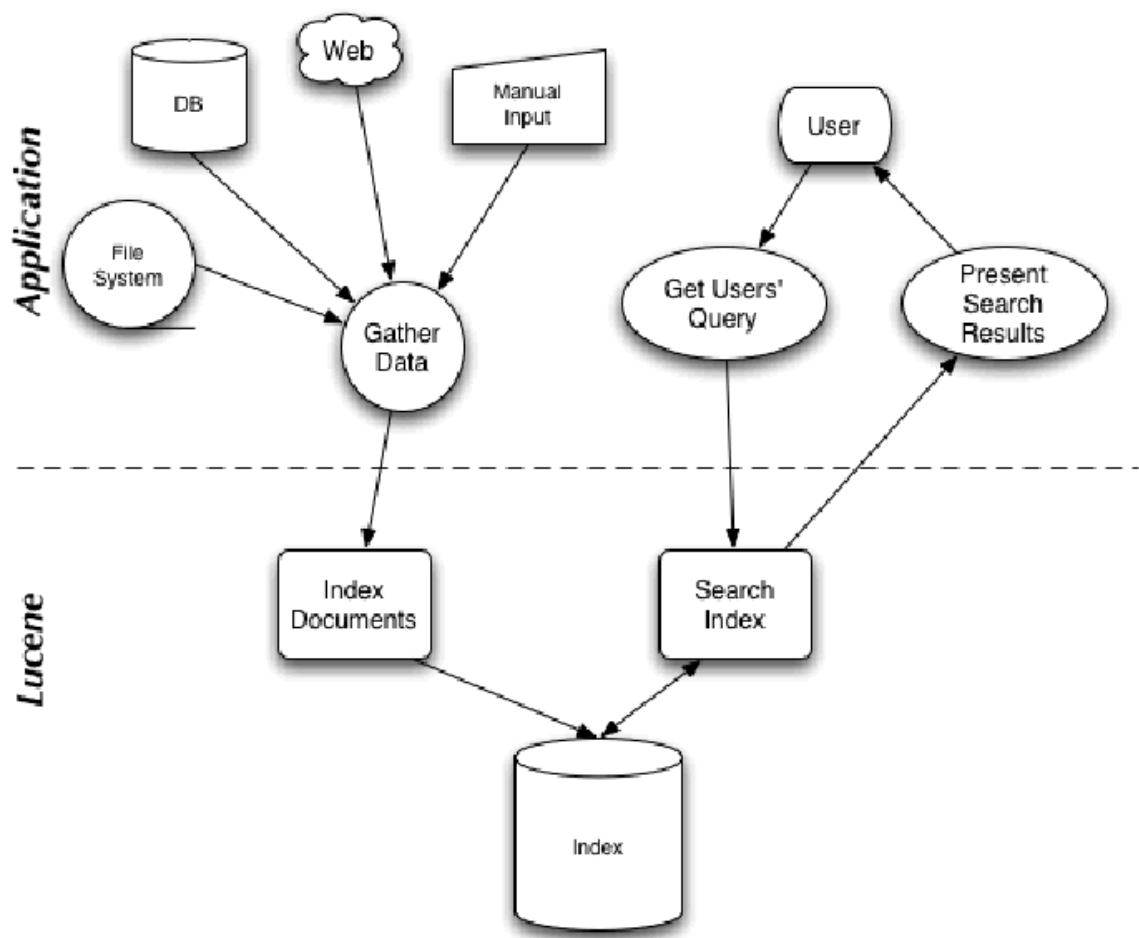


Figura 10: Arquitectura de Lógica de Lucene [59]

6.3.1. Indexando con Lucene

“Lucene usa diversos parser para diferentes tipos de documentos. Para un Documentos del HTML, por ejemplo, un programa de análisis del HTML hace cierto proceso previo, tal como filtración de las etiquetas del HTML y así sucesivamente. El programa de análisis del HTML hace salir el contenido del texto, y entonces el analizador de Lucene extrae símbolo y la información relacionada, tal como frecuencia simbólica, del contenido del texto. El analizador de Lucene entonces escribe el símbolo y la información relacionada en los archivos del índice de Lucene”. [59]

En la [Figura 11] se muestra la arquitectura de Lucene para la indexación del documento.

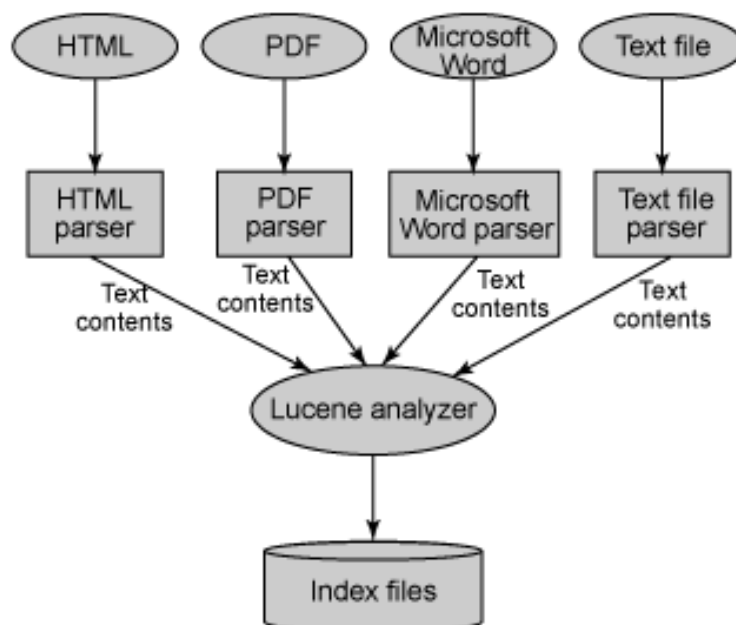


Figura 11: Arquitectura Lucene para la indexación. [59]

6.3.2. Resumen de clases que están implementadas en el Api de Lucene

- **IndexWriter:** crea el índice y agrega documentos a éste.
- **Directory:** la ubicación del índice, se puede tener un FSDirectory o un RAMDirectory, dependiendo si se desea tener dicho índice en disco o en memoria.
- **Analyzer:** proceso que analiza y limpia el texto (stop words, stemming) antes de que sea indexado.
- **Document:** una colección de Fields.
- **Field:** pareja de un nombre y valor, con que sirven para identificar las distintas partes de un documento; estas parejas pueden ser tan simples como "id" y "234" o bien tan complejas como texto completo "contenido" y "Había una vez...."

- **IndexSearcher:** Abre y emplea el índice en read-only para realizar las búsquedas.
- **Term:** unidad básica de búsqueda (son creados por el indexer)
- **Query:** distintas subclases para crear queries, BooleanQuery, PhraseQuery, RangeQuery, etc.
- **TermQuery:** un Term asociado a un query el cual puede poseer un nivel de "boost" para darle mayor peso sobre otros términos.
- **Hits:** contenedor de apuntadores hacia los resultados ranqueados de una búsqueda.

6.3.3. Servidor "Seekafire"

"Seekafire" es un servicio Windows que indexa documentos en diferentes formatos (DOC, PDF, XLS, PPT, RTF, HTML, TXT, XML y otros formatos) trabajando en segundo plano, sin que se note el tiempo de indexación, está basada en DotLucene, que es la versión en .NET del popular Lucene.

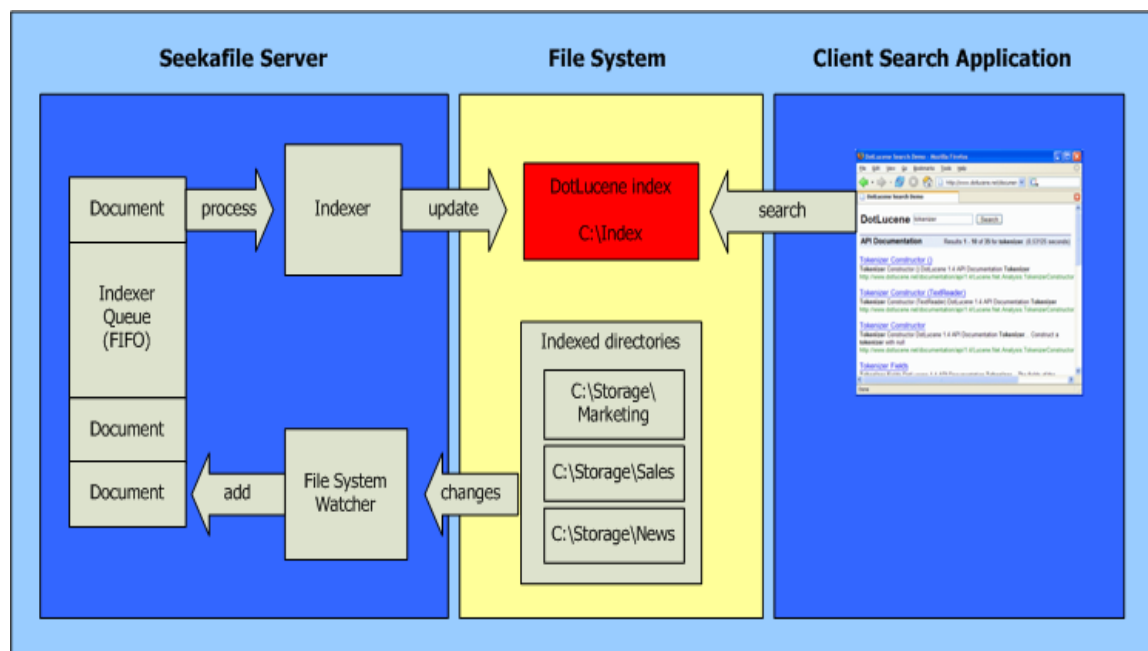


Figura 12: La descripción de la arquitectura [71].

La arquitectura es de índice céntrica. Utiliza el índice para comunicarse con los usos de la búsqueda del cliente. El índice es bastante flexible permitir lo siguiente:

- Es posible buscar el índice mientras que el servidor de Seekfile lo está modificando.
- Los cambios son visibles inmediatamente a todos los clientes.
- Las únicas informaciones que los clientes necesitan saber son: los campos, la localización del índice y del documento de DotLucene disponibles.
- El índice es compatible con la versión de Java, se puede tener acceso de un cliente de Java

Las características de Seekfile son:

Rapidez de búsqueda increíble

- Utiliza como motor [dotLucene](#) consiguiendo la mayor rapidez de búsqueda posible en el mundo .NET.
- Reconoce cualquier cambio en un segundo.

Flexibilidad

- El índice es almacenado en el formato de DotLucene/Lucene 1.3+.
- El índice puede ser utilizado directamente desde otras aplicaciones (se pueden realizar búsquedas mientras está indexando).
- El índice puede ser utilizado directamente desde aplicaciones cliente creadas con ASP.NET, Windows Forms o Java.
- El kit de instalación incluye una aplicación ASP.NET de búsqueda.

Soporte para los formatos de ficheros más usuales

Lleva incluido soporte para los siguientes formatos de fichero:

Microsoft PowerPoint (PPT)

Microsoft Word (DOC)

Microsoft Excel (XLS)

HTML (HTM/HTML)

Ficheros de texto (TXT)

Extensibilidad utilizando IFilter

Soporta [IFilter interface](#) para indexar otras extensiones de ficheros:

Adobe Acrobat (PDF)

Microsoft Visio (VSD)

Rich Text Format (RTF)

XML

etc...

Extensible utilizando *plugins* de usuario

Admite [custom plugins](#) escritos en .NET.

Mantenimiento cero

El indexador trabaja en segundo plano como un servicio Windows.

Puede especificar en el fichero de configuración las carpetas que deben ser vigiladas por si cambian.

El indexador trabaja en segundo plano con baja prioridad (no ralentiza el resto de operaciones).

Corre en Windows 2000/XP/2003 con *.NET Framework 1.1*.

Capítulo VII: Metodología Musuj Quipu

7.1. Introducción

La necesidad de preservar el contenido digital e incrementar la efectividad y velocidad en el proceso de digitalización en los servicios de búsqueda y recuperación de información, traen consigo la necesidad de crear sistemas que usen los estándares internacionales más difundidas (XML, HTML, TEI, DC, RDF) establecidas por los diversos organismos de investigación.

“MUSUJ QUIPU” es una metodología planteada para el proceso de generación, almacenamiento y publicación de los libros en la biblioteca digital Pedro Zulen (UNMSM), permitiendo preservar sus contenidos, con servicio de búsqueda y recuperación de información.

El prototipo desarrollado, contiene interfaces necesarios para el ingreso de información, que permita la generación de los documentos necesarios con el formato XML y el estándar TEI para estructurar el contenido del libro, que servirá como base de datos de archivos; para el almacenamiento del metadato se hace uso de un manejador de base de datos y los atributos se crearón de acuerdo al estándar DC. El indexado es automático a través del servidor “Seekafire”.

El sistema consta de dos etapas, la primera se creó para el ingreso del metadato, procesamiento, estructuración del libro digital y obtener un

formato estandarizado y para la segunda etapa se implementó un sistema de recuperación de información, el cual tiene el módulo de búsqueda y recuperación de los libros, con una “query string”, en todos los libros que se ha obtenido e indexado en la primera etapa.

7.2. Descripción de la Metodología Musuj Quipu

El objetivo de la metodología Musuj Quipu es: Preservar, buscar, recuperar y publicar contenidos de libros digitales. Para preservar contenidos digitales, los archivos son almacenados, estructurados y procesados de acuerdo a los estándares TEI, DC, XML, XSLT y HTML. La metodología “Musuj Quipu” consiste en los siguientes procesos:

- Proceso de marcación del documento.
- Estructuración de carpetas.
- Uso del metadato para la descripción del libro y el intercambio.
- Proceso de conversión de archivos original en formato Word a un documento XML con el estandar TEI.
- Proceso de Indexación del contenido y del metado.
- Proceso de búsqueda y recuperación del contenido almacenado del acervo bibliografico que se tiene indexado.

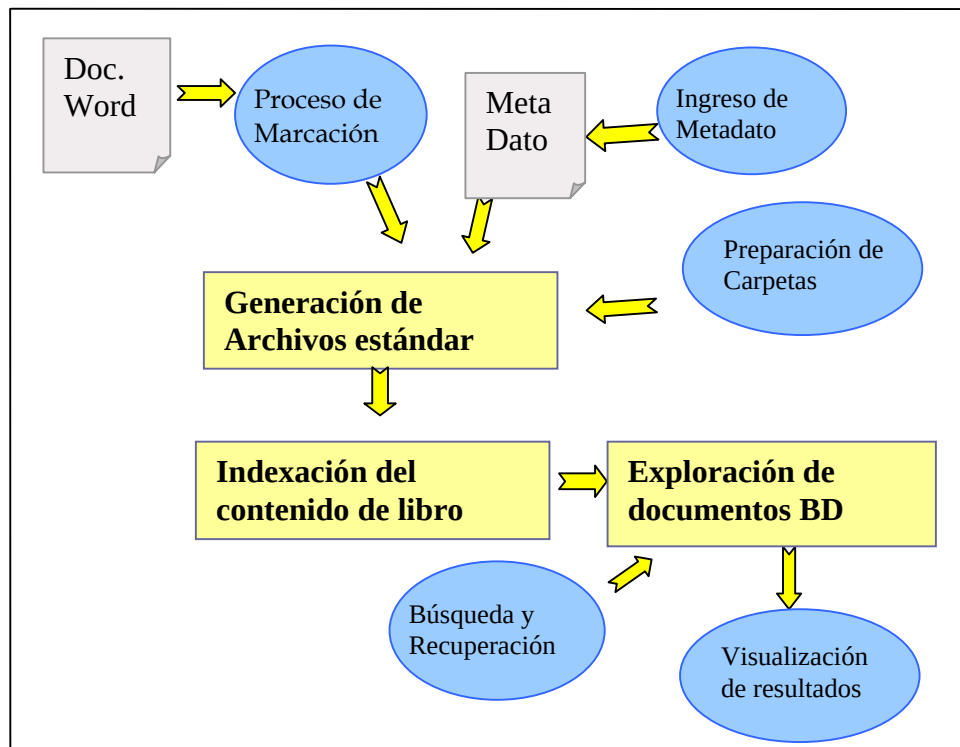


Figura 13: Procesos de Metodología Musuj Quipu [Fuente Propia].

7.2.1. Marcación del Documento

Para reconocer las diversas partes de un documento digital que se encuentra en formato de Microsoft Word, se ha creado una plantilla con campos personalizados de acuerdo al formato de un libro.

En el proceso de marcación se da formato estandarizado al documento Word de un libro, de acuerdo a las marcas creadas. Los elementos usados para el proceso de marcación se detallan a continuación.

Plantilla de marcación del documento. Es el documento creado especialmente con el formato del libro en el cual se ha dividido en partes, de acuerdo a la sugerencia del bibliotecólogo son los siguientes:

Portada. Se encuentran las marcas pertenecientes a las características que se observan en la portada de un libro y son:

- Título-Estilo (marca que representa el título principal del libro)
- Front-Subítulo (Para el subtítulo de la portada).
- Front-AutorPersonas (para marcar el autor de la obra)
- Front-Autor Secundario (para marcar el autor secundario de la obra)
- Edición (para marcar la edición de la obra)
- Editora – Estilo (para marcar la editorial de la obra)
- Ciudad Estilo (para marcar la ciudad de publicación de la obra)
- Fecha de publicación (para marcar la fecha de publicación de la obra)

Página Preliminar. Se encuentra las marcas pertenecientes a la página preliminar del libro y son:

- Dedicatoria (Para marcar la dedicatoria del libro)
- Página preliminar (Para marcar el preliminar del libro)

Cuerpo del libro. Se encuentra las marcas pertenecientes al cuerpo del libro y son:

- Partes del libro (Para marcar las partes del libro, solo se usa si el libro está dividido en partes).
- Título 1 (Para marcar el Título principal del libro).
- Título 2 (Para marcar el Título secundario del libro).
- Los demás subtítulos solo se deben marcar en negrita, para diferenciar que son subtítulos.

Citas textuales. Se encuentra las marcas pertenecientes a las citas textuales como las referencias que se encuentran en el libro y son:

- Cita Inserta del Texto (marca para los párrafos del libro)
- Cita en bloque (marca para el título principal del libro)
- Citas en verso (marca para el título secundario del libro).

Ilustraciones. Se encuentra las marcas pertenecientes al cuerpo del libro y son:

- Figura (marca para las figuras que se encuentran en el libro)

Bibliografía. Se encuentra las marcas pertenecientes a la bibliografía del libro y son:

- Referencias bibliográficas (para marcar todas las referencias bibliográficas del libro)

Texto. Sirve para marcar todo el texto (contenido) pertenecientes al cuerpo del libro.

Espacio Estilo. Sirve para marcar los espacios en blanco.

La plantilla se debe referenciar desde el Microsoft Word y luego marcar el documento tal como se ve en la [Figura 14], se debe respetar los espacios en blanco marcándolos como espacio en blanco.

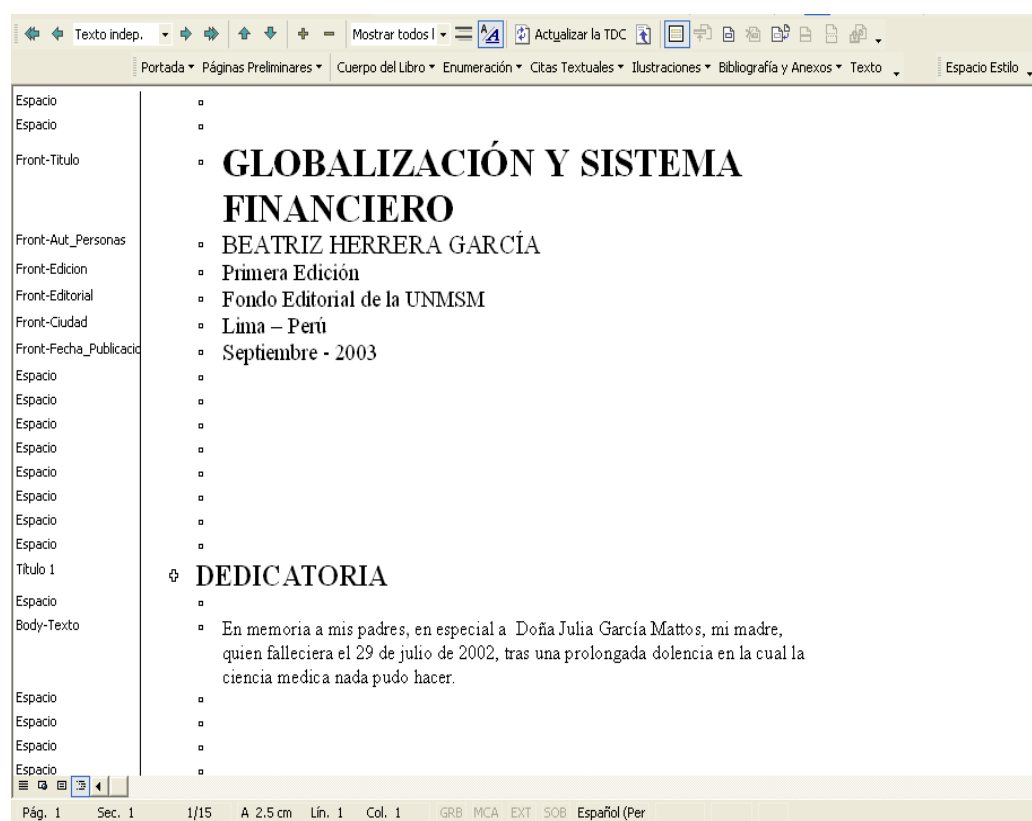


Figura 14: Documento marcado en Microsoft Word

Luego de marcar todo el libro, se deberá guardar el documento en una ruta. Posteriormente el siguiente proceso es abrir el documento xx.doc ya marcado con el Openoffice y guardarlo en la ruta de trabajo, en la carpeta recursos.

7.2.2. Estructura de Carpetas

Es necesario respetar la estructura de carpetas propuesta y especialmente la carpeta raíz, para hacerlo correctamente se debe seguir los siguientes pasos:

Paso 1. La carpeta principal donde se va almacenar todos los libros en formato digital, se debe nombrar “DigitalBook”, seguida por otro nivel de carpetas tituladas el nombre de la institución que se labora (ejemplo: SISBIB) y la carpeta “Plantilla de diseño” va contener todos los diseños

generales creados para todos los libros, el siguiente nivel de carpeta debe nombrarse con el acrónimo del autor y en el último nivel se debe tener una carpeta “Recurso” el cual va contener los archivos iniciales par poder luego procesar el libro. Vea la [Figura15].

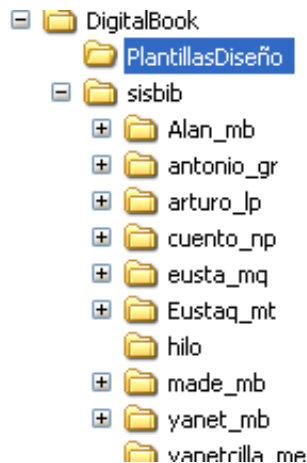


Figura 15: Estructura de las carpetas

Paso 2. La carpeta donde se va trabajar, se debe crear por cada libro, con el acrónimo y/o abreviatura del libro, colocado como identificador en el metadato, en caso que el libro tenga edición se debe colocar esta a partir de la segunda edición. El formato es como se describe a continuación:

$$\underbrace{(\text{Nombre} + _ + \text{iniciales de apellidos}) + _ + \text{Edición}}_{\text{Constante [Abreviatura]}} \quad \underbrace{\hspace{1cm}}_{\text{[variable]}}$$

Por ejemplo veamos lo siguiente:

Título del libro: Análisis de diseño y sistemas

Autor: López Hinostraza Carlos → “carlos_lh”

Identificador del libro

Análisis de diseño y sistemas (Tercera edición) → “carlos_lh_3”

Identificador del libro

Paso 3. Dentro de la carpeta antes mencionada (Paso 2) se deberá colocar una carpeta con el nombre: “Recursos” el cual contendrá inicialmente los siguientes archivos:

- o Word marcado
- o Archivo en openoffice
- o El office.dtd

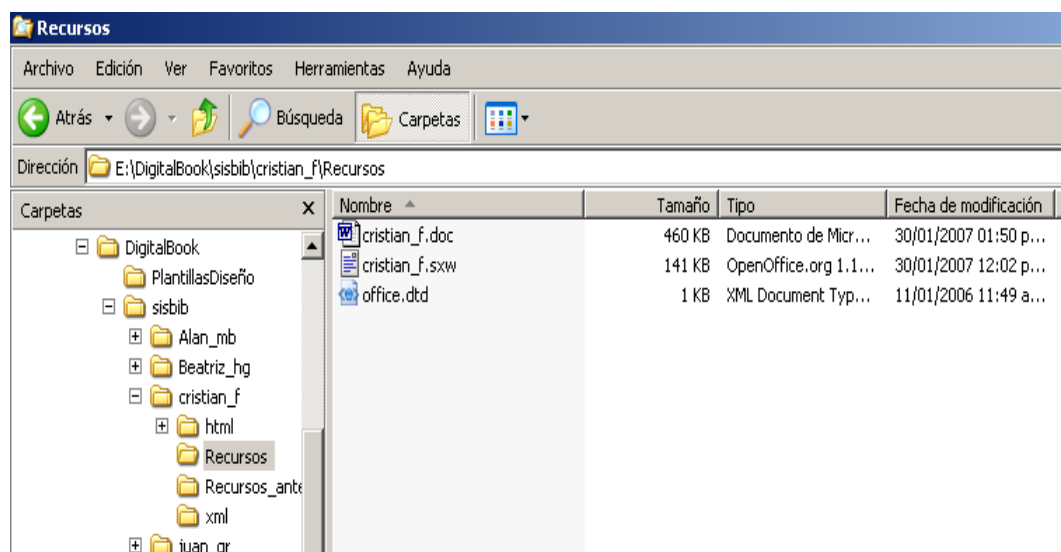


Figura 16: Contenido inicial de la carpeta “Recursos”

Paso 4. Finalmente el sistema creará las carpetas y archivos que se necesiten con sus respectivos contenidos, las carpetas que se crean son:

Xml: El cual contendrá todos los archivos xml que se obtuvo del libro, uno por capítulo.

Html: Carpeta que contiene todos los archivos en el formato del html que son las que se obtuvieron de los XML.

Las carpetas finalmente quedarán estructuradas como se observa en la [Figura 17].

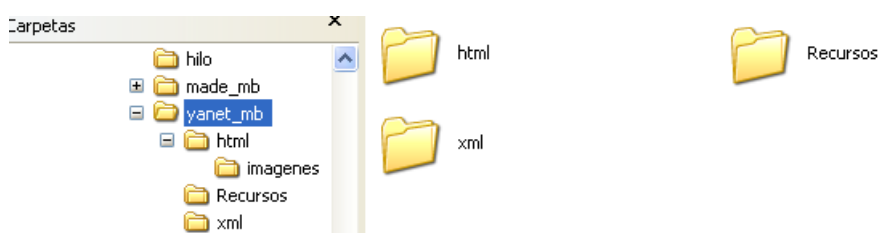


Figura 17: Carpetas creadas de manera automática

7.2.3. Uso de Metadato

El metadato es usado para describir el libro y optimizar la recuperación, los campos definidos del metadato están de acuerdo al estándar Dublin Core, en este proceso se debe ingresar los campos definidos para esta metodología, son los siguientes:

<dc:creador> : Ingresar el autor del libro, este campo puede repetirse.

<dc:title> : Se ingresa el título del libro.

<dc:subject> : Se ingresa la materia al cual pertenece el libro, como por ejemplo lenguaje, informatica, etc.

<dc:date>: Fecha de publicación del libro como por ejemplo 2006-10-25.

<dc:language>: Ingresar el idioma en el cual se publicó la obra.

<dc:description>: Ingresar el resumen del libro.

<dc:publisher> : Institución que encarga de publicar la obra digital.

<dc:identifier>: Identificador asignado al libro.

7.2.4. Proceso de conversión de Archivos

Se procesan los archivos que se ingresaron en los anteriores procesos como el documento Word marcado, el metadato y las carpetas estructuradas, se procede a crear el archivo del libro con el formato estándar TEI. Se debe reconocer las marcas ingresadas al principio y luego asignar los campos que corresponden de acuerdo al estándar TEI. Una vez obtenida el documento en formato abierto XML y estandarizado

se debe convertir el archivo en un formato legible o visualizable, para este trabajo se ha elegido el formato HTML.

7.2.5. Proceso de Indexación del contenido y del metado

Pasamos a detallar la segunda parte del proyecto, para el proceso de recuperación de información inicialmente se debe indexar el archivo al motor de búsqueda Lucene, en este proceso se realiza la indexación de los archivos generados en el proceso de generación y transformación del contenido del libro, se indexa el documento que está en formato html y el metadato ingresado es usado para describir y obtener de ella el identificador asignado a cada libro.

7.2.6. Proceso de búsqueda y recuperación

En este proceso se recupera los archivos que se han generado e indexado adecuadamente con el sistema “Musuj Quipu”, para la recuperación de información se debe usar el motor de búsqueda elegida, que es el Lucene.net y visualizar los resultados mediante un interfaz amigable al usuario.

7.3. Análisis y diseño del sistema Musuj Quipu

Para el análisis y diseño del sistema se usó el lenguaje unificado de modelamiento UML.

EL sistema está compuesto por dos partes y son:

- El ingreso y la generación del documento con el formato estándar TEI y DC.
- La indexación, búsqueda y recuperación de la información.

7.3.1. Actores del sistema

El sistema Musuj Quipu cuenta con tres actores:



Actor Bibliotecólogo: Es el usuario especializado, encargado de ingresar nuevos libros al sistema, debe ingresar el metadato, realizar la marcación del libro, ingresar autores, modificar los datos ingresados, crear la estructura de los archivos en una ruta específica y finalmente ejecutar el evento de generación del libro.

Actor Usuario: Representa a las personas que realizarán la operación de consulta, realizará búsquedas de acuerdo a los parámetros ingresados y visualizará los resultados y el contenido del libro.

Actor Lucene.net: Representa al sistema especializado en el almacenamiento y recuperación de información, es el motor elegido para este sistema.

7.3.2. Diagrama de casos de uso general

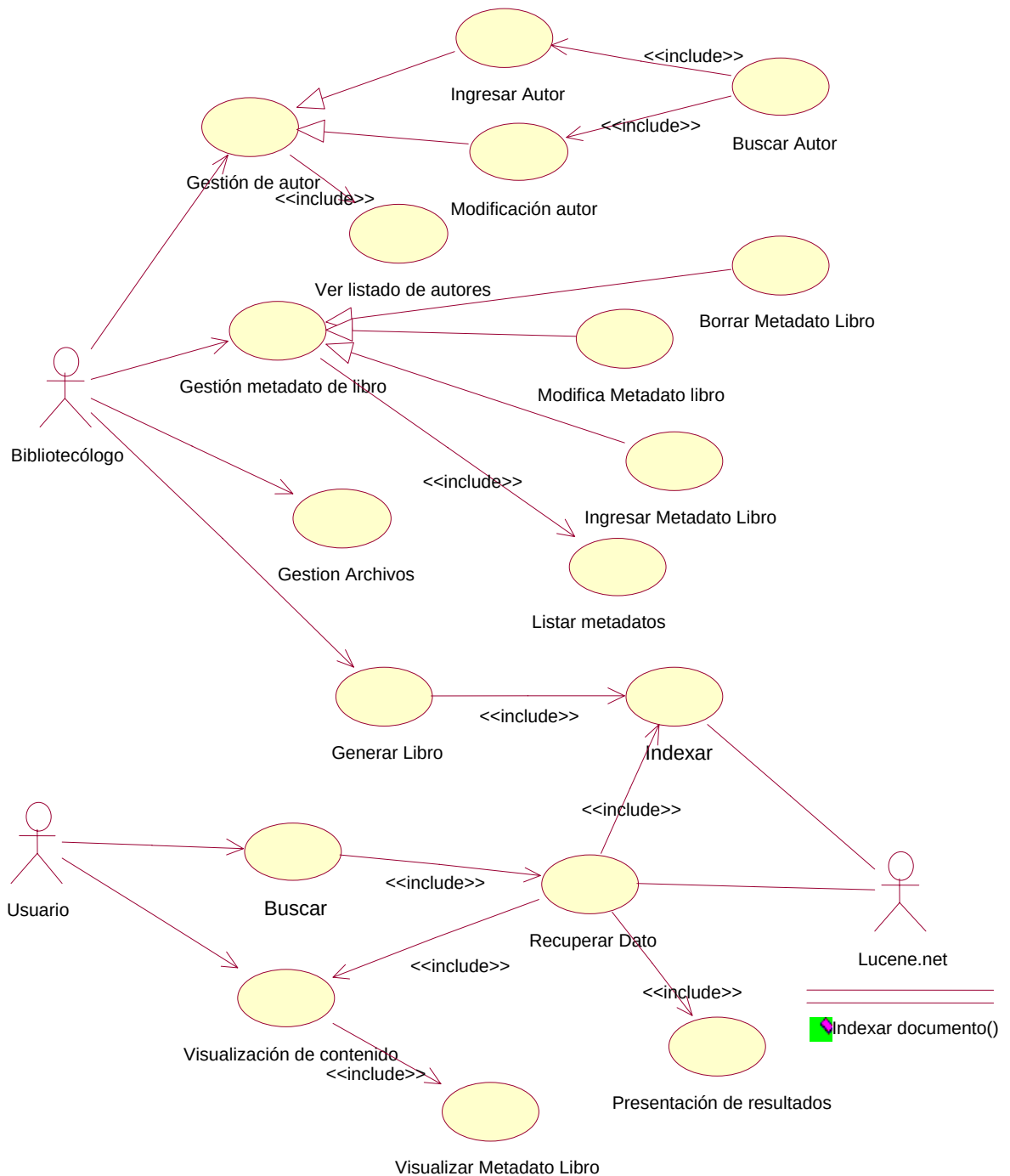


Diagrama 1: Diagrama de caso de uso general

7.3.3. Modelo de datos para el Contenido del Libro

Para el almacenamiento y estructuración del contenido del libro se ha usado el DTD TEI, en ella se define las reglas que se obtuvo de la siguiente manera.

“La operatividad de TEI se ha establecido entorno a cuatro comités que comparten la responsabilidad de elaborar las directrices.

1. El Comité de Documentación de Textos se encarga de definir las etiquetas para identificar los textos (procedencia, ubicación, clase, categoría, etc.).
2. El Comité de Representación Textual se ocupa de describir física y lógicamente los textos. La descripción lógica abarca cuestiones como su estructura (capítulos, secciones, etc.), la tipografía, la maquetación, las notas, apéndices y referencias diversas, etc.
3. El Comité de Análisis e Interpretación de Textos trata el desarrollo de etiquetas que permiten la descripción lingüística y literaria del texto, así como cuestiones de intertextualidad, indización, etc.
4. El Comité de Cuestiones Metalingüísticas se ocupa de los problemas técnicos de la sintaxis utilizada en la etiquetación.

Las cuestiones dirimidas por estos comités de TEI dan idea del grado de complejidad y precisión al que se quiere llegar. La aplicación de estas directrices permite la transformación radical en la difusión y acceso al conocimiento.”[63]

El contenido de libros es almacenado en una base de datos no estructurado (en archivo XML).

7.3.4. Modelo de datos del Metadato

Se ha realizado el modelado de datos para almacenar el metadato del libro. Los atributos considerados en la entidad principal (Cyb_Libro) son los elementos más importantes descritos en la metodología Dublín Core tal como se observa en la [Figura 18], éstas se detallaron en el punto [5.3.2] de este trabajo. Las entidades complementarias son “Cyb_Creator”, en donde se almacena los atributos de los autores (como nombres, apellidos, fecha de nacimiento, biografía, etc.), la entidad Cyb_materias, contiene las diversas materias al cual pertenece un libro (ejemplo: Lenguaje, química, física, matemática, etc.).

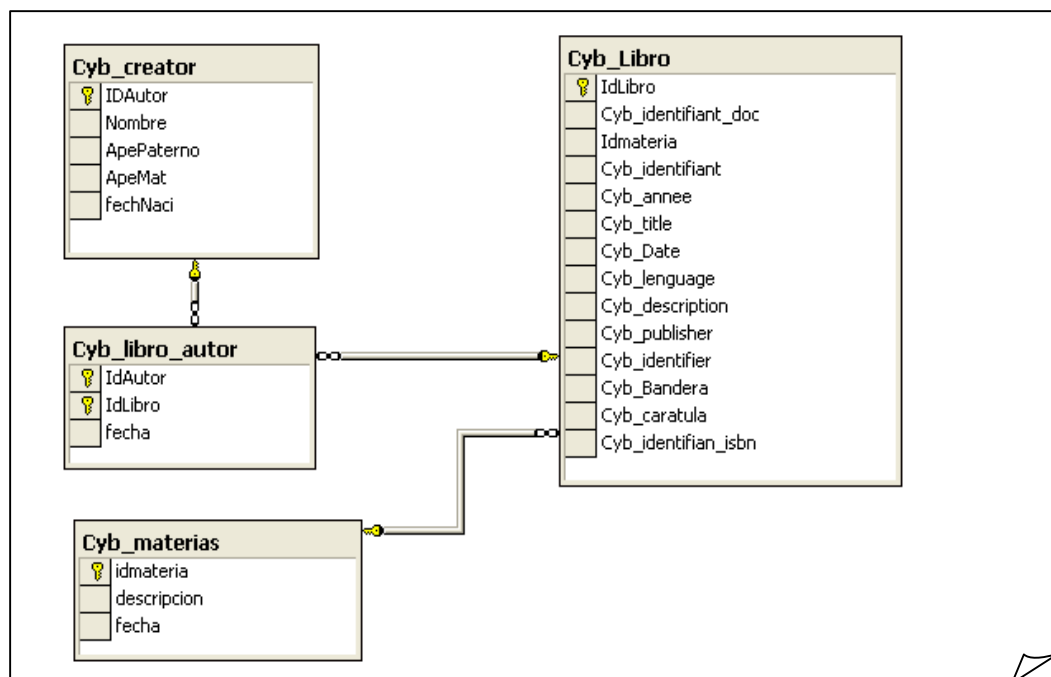


Figura 18: Diagrama entidad relación del metadato.

7.3.5. Arquitectura de Musuj Quipu UNMSM

La arquitectura del sistema Musuj Quipu UNMSM se muestra en la [Figura 19], en donde se detalla la división del sistema en dos partes bien diferenciadas.

1. La primera es el ingreso de datos a través del módulo en escritorio, permitiendo este el ingreso de dos tipos de datos:
 - El contenido del libro y el metadato del mismo.
 - Posteriormente en base a los datos ingresados se procederá a generar los documentos y realizar la indexación del contenido a la base de archivos y el metadato a la base de datos.
2. La segunda parte es el módulo que se encarga de la búsqueda y recuperación de información en base a la Query ingresada por el usuario, para ello se hace uso del índice generado por el sistema "Servidor Seekafire". En la [Figura 19] se muestra la arquitectura del buscador que es de índice-céntrica, usa el índice para comunicarse con los usos de la búsqueda del cliente. El índice es muy flexible y permite lo siguiente:
 - La búsqueda mientras se esta modificando el índice.
 - Los cambios son visibles inmediatamente.
 - El índice es compatible con el índice en java.
 - La única información que los clientes necesitan saber son los campos de la localización del índice y del documento de DotLucene disponibles.

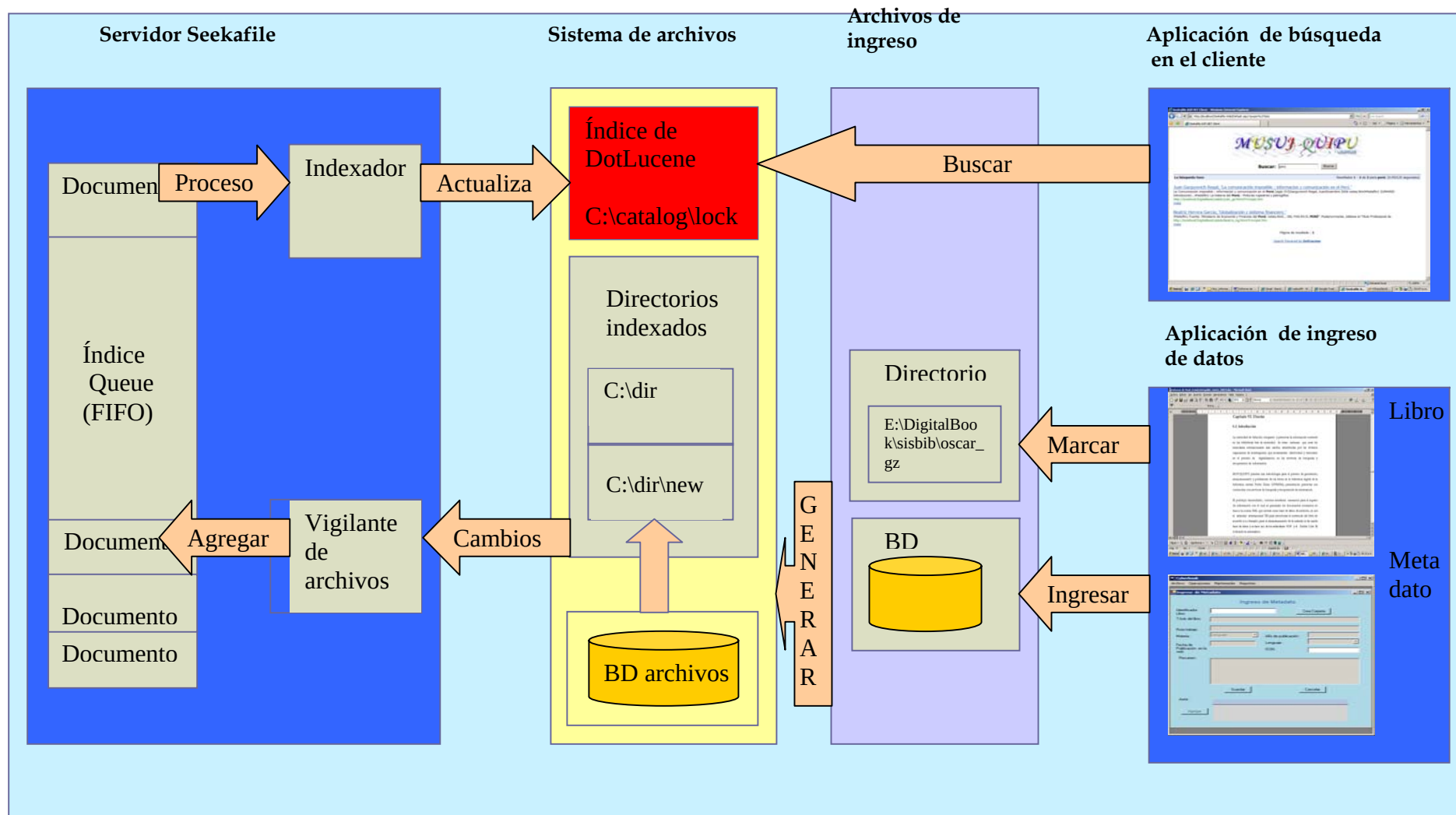


Figura 19: Visión de la arquitectura de Musuj Quipu [Fuente Propia]

7.4. Principales módulos del sistema

Para la implementación del sistema Musuj Quipu se divide en tres módulos funcionales bien diferenciados, los mismos que se muestran en la [Figura 20] y luego pasaremos a describir cada uno de ellos.

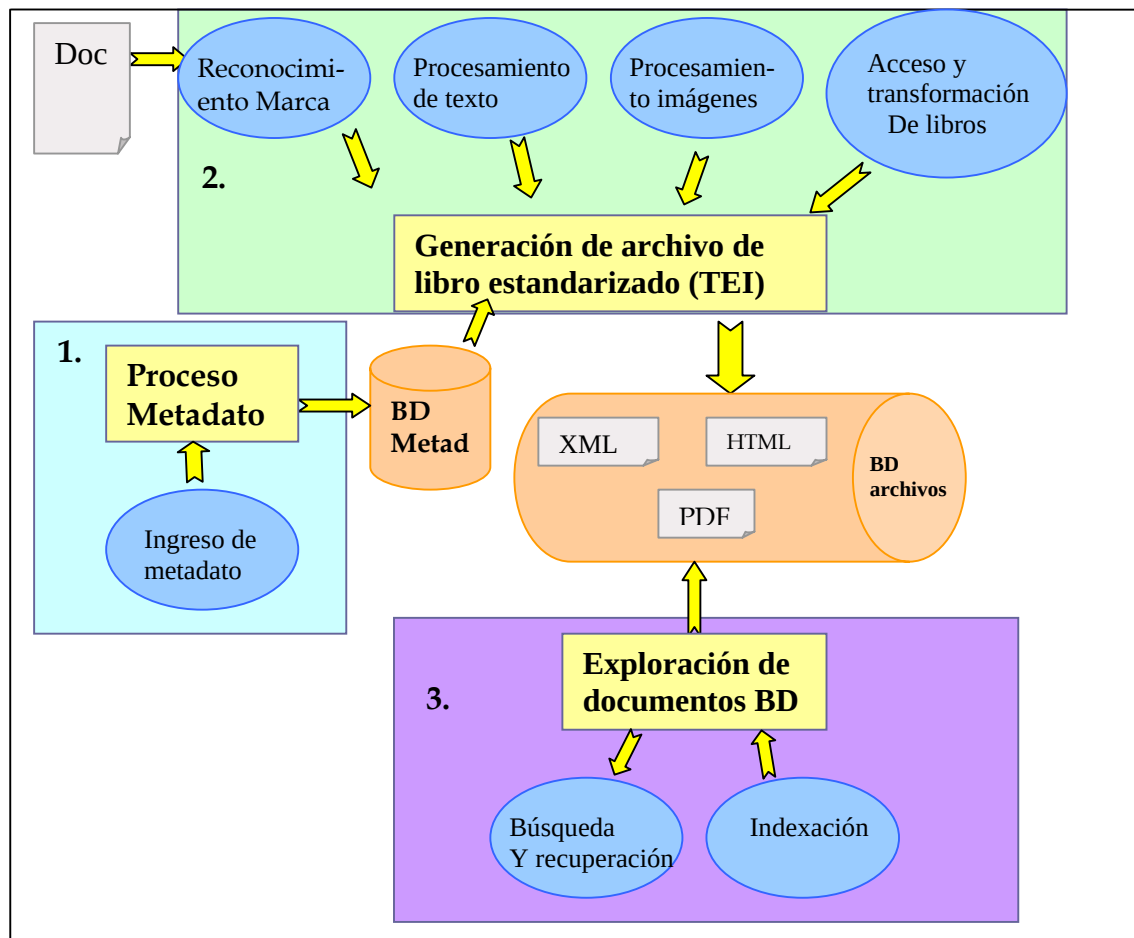


Figura 20: Módulos funcionales del sistema “Musuj Quipu”

[Fuente Propia].

1. Proceso de metadato

Es necesario agregar alguna descripción a los libros almacenados, para facilitar al usuario, se ha creado un módulo por el cual se va administrar el metadato e ingresar datos como: título, autor, institución, año publicación, etc. Esta información es adicional y está referenciado a cada libro ingresado.

La indexación del metadato se realiza a través del módulo del sistema “Musuj Quipu_ Generador”, para este caso indexación significa crear un índice en base al metadato ingresado, para ello se usa el MSQl 2000.

2. Generación de archivo del libro

Para añadir un libro a la base de datos de archivos, este pasa por cuatro etapas, se requiere que inicialmente se ingrese el libro en formato xx.doc y luego los procesos a realizar son los siguientes:

- *Reconocimiento de Marca*, proceso en el cual se dará el formato al documento Word de acuerdo a las marcas creadas en el Microsoft Word.
- *Procesamiento de Texto*, proceso interno en el cual el sistema reconoce las marcas de los textos y de acuerdo a ellas se usa los datos contenidos en cada marca para generar el documento de acuerdo al estándar TEI respetando el tipo de marca colocado y obtener un documento en XML.
- *Procesamiento de imágenes*, se reconoce las marcas especiales de las imágenes que contiene el libro y se reescribe las marcas de acuerdo al estándar TEI. Las imágenes son procesadas, para su disponibilidad.
- *Transformación del documento*, proceso en el cual se convierte el documento XML generado de acuerdo al estándar TEI, al formato HMTL, para ello se diseñó XSLs.

3. Exploración de los documentos

El usuario final debe ser capaz de realizar búsquedas sobre temas específicos así como consultar a los libros almacenados. Los resultados

serán obtenidos de acuerdo a un sistema especializado (Lucene.net) en almacenamiento y recuperación de información.

En el siguiente capítulo se describe los detalles técnicos de la implementación de las funcionalidades descritas.

7.5. Implementación del sistema Musuj Quipu

Para demostrar la viabilidad de la arquitectura descrita en base a la metodología propuesta, se ha implementado el prototipo denominado “Sistema Musuj Quipu”, específicamente para el procesamiento de los libros digitales.

A continuación se describe la implementación de la metodología Musuj Quipu, para ello se han usado diversas herramientas que se detallan en los siguientes párrafos.

7.5.1. Selección de tecnologías

Para seleccionar las tecnologías que se usaron, se tuvo que tener en cuenta varios factores como las efectividad, disponibilidad y eficiencia para el caso requerido. La selección del lenguaje para el tratamiento del contenido de los libros fue un factor decisivo tanto para la etapa del diseño de los interfaces como para la realización del módulo generador, por ello después de analizar esos factores se decidió en usar las siguientes tecnologías:

- El XML se usó para el almacenamiento del contenido del libro, por ser un metalenguaje de marcado, por su facilidad, y por las características detalladas en los capítulos anteriores.

- XSL, XML permite crear diferentes interfaces a través del uso del XSLT que realiza la transformación tomando como hoja de estilo el XSL.
- HTML, formato ampliamente usado en el Internet y se ha aprovechado su simplicidad para mostrar el contenido del libro.
- Para la programación se usó en lenguaje orientado a objetos Microsoft. NET 2003.
- Microsoft SQL 2000, como gestor de base de datos para el almacenamiento de los metadatos.
- Open Office Versión 1.1.4, procesador de texto libre, usado para el proceso de transformación.
- Microsoft Word 2000, se usa para el marcado del contenido del libro, debido a que actualmente es el procesador de texto más usado, y es el requerimiento de la bibliotecóloga del área de biblioteca digital Pedro Zulen.
- DTD TEI, se usó para crear el archivo XML con los campos del DTD estándar para preservar el contenido del libro.
- Oxygen XML Editor, Herramienta usado para crear los XSLT y revisar los XMLs.

7.5.2. Especificaciones Técnicas

Para el funcionamiento correcto del sistema, los clientes deben tener en cuenta las siguientes especificaciones técnicas:

- Microsoft Word XP/ Word2000
- Open Office Version 1.1.4
- Sistema operativo Windows XP
- Tener instalado el Framework del .NET 2003
- Maquina Pentium III.

- Explorer 5.0
- Sql2000

7.5.3. Desarrollo del Sistema Musuj Quipu

Como se mencionó en el capítulo anterior el sistema Musuj Quipu consta de dos subsistemas bien diferenciados, las mismas que se han desarrollado de manera complementaria y son:

- Aplicación de ingreso de datos denominado “Sistema Generador”.
- Aplicación de búsqueda de datos en el cliente “Musuj Quipu UNMSM”.

7.5.3.1. Componentes del Sistema Generador

La aplicación de ingreso de datos, es el sistema por el cual se ingresa los datos necesarios para el procesamiento y obtención del libro digital de acuerdo a los estándares propuestos en los capítulos anteriores, sus principales módulos son:

- **Ingreso de metadato**, módulo por el cual se ingresa el metadato del libro a procesar, a través de un interfaz, los campos definidos para este prototipo están de acuerdo al estándar Dublin Core.
- **Ingreso de autor**, se ingresa el autor ó autores del libro que se está procesando, solo en el caso que todavía no estén registrados en el sistema.

- **Listado de libros ingresados**, módulo por el cual se puede observar los detalles de los libros ingresados a fin de tener una mejor administración de ellos.
- **Ingreso de materias**, mediante este módulo se ingresan las materias a los cuales pertenecen los libros que se van a procesar.
- **Módulo Generador**, el generador es el módulo de producción que organiza el contenido del documento digital que inicialmente está en formato xxx.doc marcado y se obtienen los documentos XML estructurado con el estándar TEI y estos son transformados de acuerdo a las hojas de estilo XSL diseñados previamente y finalmente se obtiene el documento con el formato HTML. Los procesos descritos en el párrafo son para cada uno de los libros que se quiera agregar a nuestra colección de materiales publicados, en la [Figura 21] se detalla el flujo de trabajo y la secuencia del proceso interno del programa.

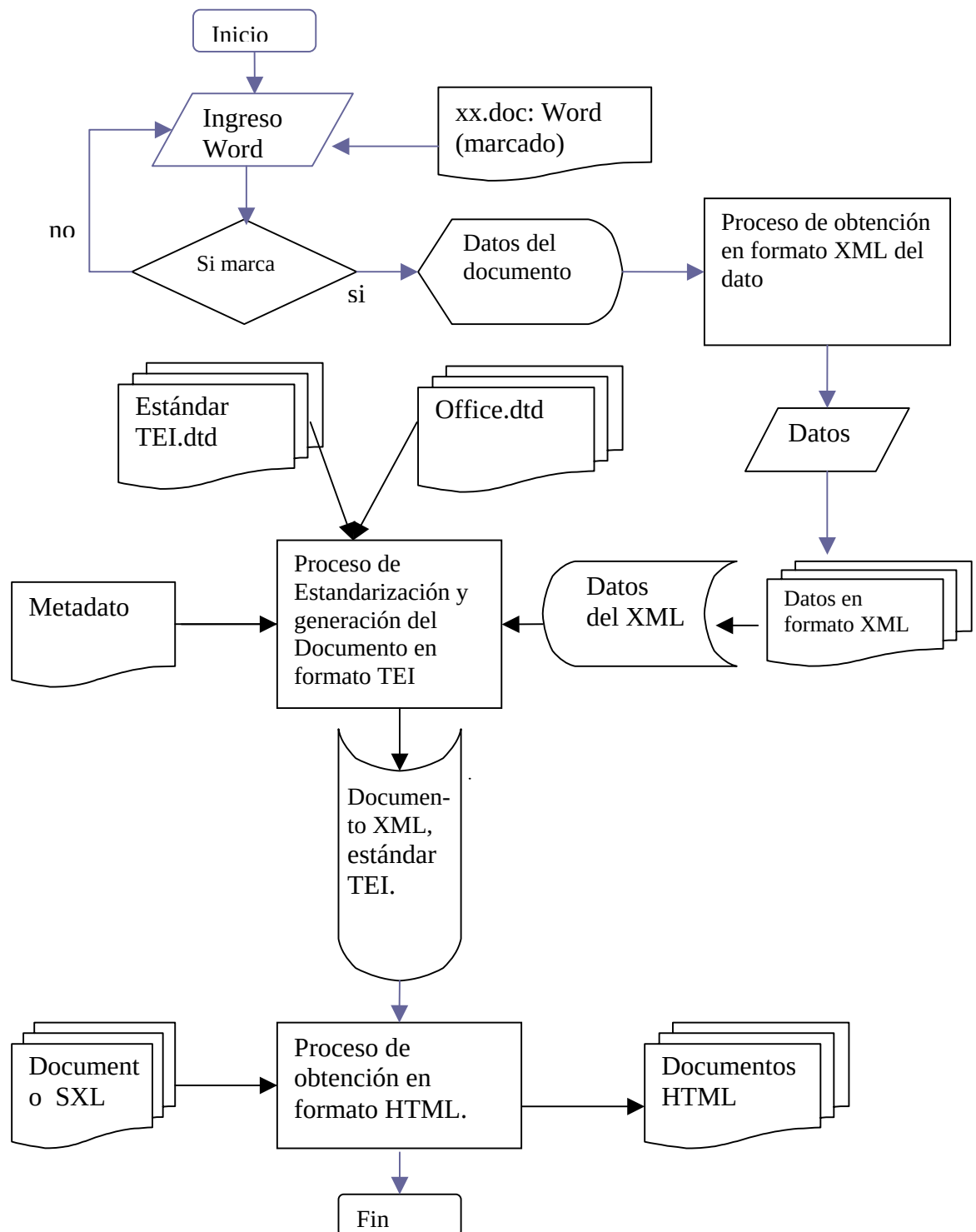


Figura 21: Diagrama de Flujo del Sistema Generador

El proceso de obtención del documento en formato XML.- En este proceso se obtiene los datos en formato XML, estructura de Microsoft Word, listo para ser analizado en el siguiente proceso.

El proceso de generación del documento de acuerdo al estándar del TEI.- En este proceso, se obtiene el documento XML con las marcas de la norma TEI, se procesan todas las características del contenido del documento tales como: negritas, títulos, los subtítulos, el tratamiento de imágenes, la bibliografías y capítulos, en este mismo proceso se enlaza con el identificador asignada en el metadato al libro.

Proceso de obtención en formato HTML.- Este es el último proceso en el cual se ingresan los archivos XSL previamente diseñados, y el documento estandarizado con el TEI para luego obtener como resultado, los documentos con el formato HTML con una página principal para consultar todas las partes del libro digital procesado.

7.5.3.2. Formato del documento XML con el Estándar TEI

A modo de ejemplo se tiene el siguiente documento que se ha obtenido de la aplicación “Sistema Generador”, en el cual se observa que el documento se divide en partes, de acuerdo al estándar TEI y estas son : front, body y back. Los elementos usados en el documento XML están definidos también de acuerdo al DTD del TEI, por ejemplo el elemento <docAuthor> en el TEI está definido para guardar el autor, dentro de los <p> deben ir los párrafos, y <div id="TH.2" n="2" type="capitulo"> está separando cada capítulo del libro, por ello es secuencial el ID que se le asignado.

```

<?xml version="1.0" encoding="UTF-8" ?>
<TEI.2 id="sisbib_GB">
<teiHeader></teiHeader>
<text>
<front>
  <titlePage>
    <docTitle>
      <titlePart type="main">
        <emph rend="bold">GLOBALIZACIÓN Y SISTEMA FINANCIERO</emph>
      </titlePart>
    </docTitle>
    <byline>
      <docAuthor>BEATRIZ HERRERA GARCÍA</docAuthor>
    </byline>
    <docEdition>
      <emph rend="bold">Primera Edición</emph>
    </docEdition>
    <docImprint>
      <emph rend="bold">Fondo Editorial de la UNMSM</emph>
    </docImprint>
    <docDate>
      <emph rend="bold"> Lima - Perú <lb /> Septiembre
        - 2003 </emph>
    </docDate>
  </titlePage>
</front>
<body>
<div id="TH.1" n="1" type="capitulo">
<head>DEDICATORIA</head>
<p>En memoria a mis padres, en especial a Doña Julia García
Mattos, mi madre, quien falleciera el 29 de julio de 2002, tras
una prolongada dolencia en la cual la ciencia medica nada pudo
hacer.</p>
</div>
<div id="TH.2" n="2" type="capitulo">
<head>PRESENTACIÓN</head>
<p>El presente libro es el resultado de investigaciones conjuntas
realizadas por el Instituto de Investigación en Ciencias
Financieras y Contables de esta cuatricentenaria casa de
estudios. El objeto fue realizar un análisis de los problemas
financieros, especialmente sobre los aspectos de la globalización
y las crisis económicas más recientes. Es decir, estudiar cómo la
internacionalización del negocio bancario en condiciones de
deficiencias del sector, así como una inadecuada supervisión de
esas entidades han desempeñado un papel central en las crisis
financieras recientes.</p>
<p>Con posterioridad a la crisis financiera asiática, y durante el
corto período 1997-1998, se volvió evidente que existía una mayor
interdependencia entre las diferentes regiones del mundo y entre
estas economías y la de Estados Unidos. De hecho, la economía
estadounidense actuó como una locomotora, principalmente en la
segunda mitad de la década de los años 90, ya que con la
expansión de su enorme mercado interno permitió que se superara
la recesión de muchas economías en desarrollo seriamente
afectadas por problemas financieros. Sin embargo, en el caso
latinoamericano el efecto "dragón" y el efecto "vodka" tuvieron
mayor incidencia en la larga recesión iniciada en 1998.</p>
</div>
<div id="TH.3" n="3" type="capitulo">

```

```

<head>CAPÍTULO I</head>
<p>
<emph rend="bold">EL NUEVO MILENIO Y LAS CONDICIONES ECONÓMICO -
FINANCIERAS</emph>
</p>
<p/>
<p>
<emph rend="bold">Introducción</emph>
</p>
<p>
Nunca la economía mundial se había manifestado tan dependiente de
la de los Estados Unidos como en los años 2001, 2002 y 2003. En
la reciente década de los años 90 del siglo pasado la economía
norteamericana fue la principal fuente de crecimiento de la
demanda mundial,
<emph rend="super">
<note n="1" id="ftn1">
<p>En el período 1996-2000, América del Norte (Estados Unidos y
Canadá) creció a una tasa promedio anual del 4,0 %, lo que
representa el gran dinamismo de esa región.</p>
</note>
</emph>
</p>
<p />
<p>
Desde el 24 de octubre de 1929 –inicio del derrumbe de la Bolsa
de Valores de Nueva York y de la significativa caída del índice
Dow Jones– el valor de las acciones, bonos y demás títulos se
volatilizó produciendo la ruina de los inversionistas.
<emph rend="super">
<note n="2" id="ftn2">
<p>Este hecho marcó el comienzo de la depresión a escala mundial,
sus efectos negativos alcanzan hasta 1933. Según Devlin, a la
crisis de 1929 siguió una aguda baja de los precios de los
productos básicos, así como del volumen comercial. Las economías
de muchos países en desarrollo se derrumbaron bajo el peso de su
deuda externa y la consiguiente desintegración de la economía
mundial llevó a prolongados períodos de mora. En 1933 se hallaba
impago el 67% de las emisiones latinoamericanas de bonos en
dólares (ver, R. Devlin, Los Bancos Transnacionales y el
Financiamiento Externo, p. 11)</p>
</note>
</emph>
</p>

</div>
</body>
<back>
<div type="bibl" id="TH.back.1">
<head>REFERENCIAS BIBLIOGRÁFICAS</head>
<bibl>AGMON, Tamir; Robert HAWKINS y Richard LEVICH, El futuro del
sistema monetario internacional, México, D.F., Prisma,
1986.</bibl>
<bibl>Banco Interamericano de Desarrollo, Deuda pública externa de
los países de América Latina, Washington, D.C., 1978.</bibl>
<bibl>RIVAS, S., P., Teoría y Política Monetaria y Bancaria, Lima,
1997.</bibl>

```

```
<bibl>ROJAS-SUÁREZ, Liliana y Steven WEISBROAD, Banking Crises in  
Latin America: Experiences and Issues, Washington, D.C., Banco  
Interamericano de Desarrollo, 1995.</bibl>  
<bibl>ROJAS-SUAREZ, Liliana y Steven WEISBROAD, Towards an  
Effective Regulatory and Supervisory Framework for Latin America,  
Washington, D.C., Banco Interamericano de Desarrollo, 199.</bibl>  
<bibl>Superintendencia de Banca y Seguros, Información financiera  
mensual, 1998 y 1999, Lima.</bibl>  
</div>  
</back>  
</text>  
</TEI.2>
```

Este documento, luego pasa por el transformador XSLT y se obtienen los documentos HTML correspondientes. Ver el [Anexo 3] en donde se observa los diseños de las hojas de estilo XSL.

En el [Anexo 4] se presenta un recorrido por el subsistema “Aplicación de ingreso de datos, Sistema Generador” y en [Anexo 5] se encuentra el recorrido por el interfaz del Sistema Musuj Quipu _ UNMSM.

7.5.3.3. Componentes de la aplicación de búsqueda de datos en el cliente

El subsistema aplicación de búsqueda de datos en el cliente, permite brindar el servicio de recuperación de información de acuerdo a una consulta realizada, para el desarrollo de este sistema se ha usado el SRI Lucene.net a través del sistema “Seekafire.net” que es un indexador que se ha desarrollado en el lenguaje Microsoft .NET. Sus características se han detallado en el capítulo Séis de este trabajo de tesis.

Para implantar y adecuar Seekafire.net al sistema Musuj Quipu se tuvo que realizar las siguientes tareas:

- Investigar la forma de trabajo del SRI.
- Hacer las configuraciones necesarias en el sistema operativo.
- Ejecutar el servicio para que realice la indexación.

- Averiguar su arquitectura, modificarlo y adecuarlo para este caso específico, el buscador debe obtener el documento y el metadato del libro para mostrarlo en el resumen del resultado obtenido por el motor de recuperación de información.
- Configurar las carpetas necesarias en la ruta especificadas para su funcionamiento.

El buscador se desarrolló en entorno web, debido a que el sistema esta orientado al servicio de la comunidad en general, la página se muestra en la [Figura 22], donde podemos hacer las consultas a todos los libros indexados hasta la fecha, la recuperación de información es en base al contenido.



Figura 22: Portada en Web del sistema Musuj Quipu

7.6. Descripción del trabajo desarrollado

Inicialmente se empezó a investigar sobre las diversas formas de desarrollar la solución para el problema planteado en el capítulo dos de este trabajo, además de averiguar sobre las diversas tecnologías existentes tales como TEI, XML, RDF, META DATO, DC. Se realizó las siguientes acciones:

1. Se investigó sobre las funciones de las bibliotecas digitales a través de la historia, el rol que cumple en la sociedad de la información, y cuales son sus funciones actuales.
2. Se Investigó sobre las diversas formas de implementar las bibliotecas digitales, sobre las tecnologías usadas para ello, y los estándares más usadas, eligiendo las más consistentes y mas usadas.
3. Se analizó la situación actual de la biblioteca digital de la universidad, tomando los datos existentes como datos de ingreso, se planteó la metodología Musuj Quipu, se creó el diagrama de casos de uso y se diseñó el modelo de datos.
4. Se esbozó la arquitectura del sistema Musuj Quipu para empezar a desarrollar.
5. Para el desarrollo del sistema se tuvo que realizar las acciones que se detallan en los puntos siguientes.
6. **Análisis el documento Word.-** En esta parte se analizó la funcionalidad del Microsoft Word, averiguar como crear los

botones para las poder realizar las marcas, la creación de plantillas, y seguidamente se analizó el documento XML que genera el Microsoft Word y ver los diversos tipos de marcas que crea, la secuencia y la forma como usa sus marcas y las relaciones de las mismas con las marcas nuevas creadas. Finalmente crear la plantilla Word personalizada.

7. Se investigó la forma de obtener el XML del Documento Microsoft Word para poder trabajar con el contenido del mismo.
8. Se investigó las partes que tiene un libro y los formatos y/o estándares para las publicaciones físicas que cumplen, para poder crear las marcas correspondientes.
9. Se analizó el documento Word y el OpenOffice para la obtención de los datos en formato XML.
10. Investigación sobre los documentos que cumplen el formato del TEI y documentarse sobre los estándares XM, TEI.
11. Se implementó el “Sistema Generador” que se va encargar de ingresar los datos en el servidor, el cual obtener y estructurar el contenido que está en el formato xx.doc luego pasarlo al formato XML y poder organizar de acuerdo al estándar TEI, hacer el reconocimiento de todos los elementos del contenido tales como, títulos, imágenes, letras negritas, etc.
12. Investigación del XSLT, para poder convertir los documentos del formato XML al formato HTML.

13. Se construyó las plantillas de conversión para cada tipo de los documentos existentes, tales como portada, pagina principal, contenido de capítulos, pies de página y la bibliografía.
14. Se empezó a diseñar la estructura del metadato y se creó un interfaz para el ingreso de metadatos, para ello se investigó sobre la creación, uso de los metadatos y los diversos estándares que existen para ello, tales como Dublín Core (DC) y el RDF. Y finalmente con los campos del DC Se creó la base de datos para almacenar los metadatos.
15. Se investigó sobre los diversos tipos de motores de recuperación de información eligiendo la más adecuada y potente (Lucene.net). Luego este motor se modificó y adecuó al sistema Musuj Quipu para poder trabajar con el tipo de datos (indexación de HTML) que se tiene como resultado del sistema generador.
16. Finalmente se realizó las pruebas al sistema en general.

Capítulo VIII: Análisis de los resultados obtenidos y trabajos futuros

8.1. Análisis de los resultados

Para dar robustez y seguridad al prototipo del software desarrollado se han realizado diferentes pruebas, los que se narran de manera sucinta en los siguientes párrafos.

8.1.1. Consistencia de base datos

Las primeras pruebas efectuadas a la aplicación desarrollada tuvieron que ver con el acceso a la base de datos desde el cliente. Se realizó lo siguiente:

Se agregó y se borró quince metadatos de libros en la base de datos, mostrando que su acceso es seguro y consistente y que todos los metadatos asociados a cada libro son creados y borrados de manera independiente al contenido del libro.

Debido al manejo manual del código asignado a cada libro, se creó a todas las tablas sus llaves primarias, de esta manera los datos que se ingresan son consistentes, se valida en el servidor la repetición de código de libro asignado antes de ingresarlo, para evitar repetición en la creación de las mismas.

8.1.2. Análisis del tiempo de procesamiento del documento digital

Según las pruebas realizadas de la generación se ha tenido las siguientes observaciones:

1.- Se ha disminuido el tiempo de procesamiento y publicación de un libro digital de 9 horas y 30 minutos a 1: hora y 35 minutos, tiempo aproximado para un libro de 100 páginas con imágenes incluidas que se detallan en la [Figura 23] y [Figura 24]

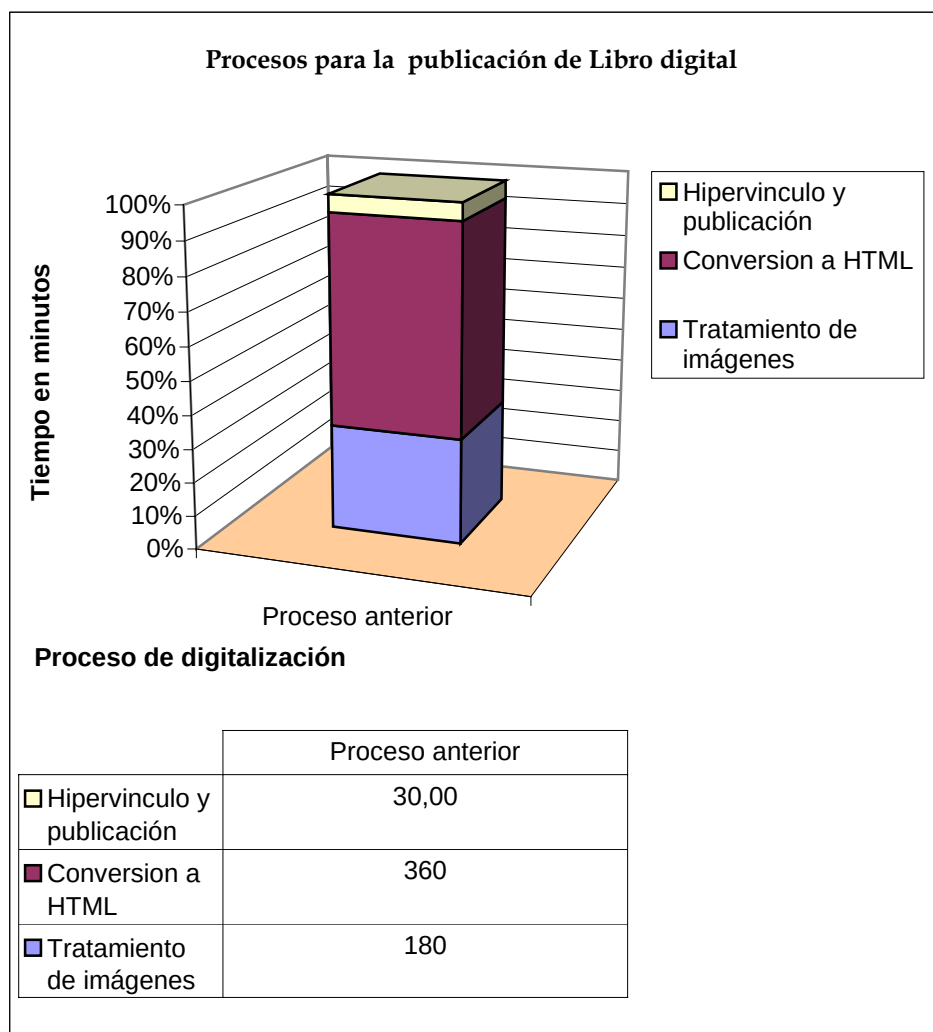


Figura 23: Procesos actuales y sus tiempos

Los tiempos de la figura están estimados en minutos, los procesos del sistema Musuj Quipu para la generación son: marcación, preparación de carpetas, generación, indexación que se observan en la [Figura 24], de igual manera se muestran en la [Figura 23] los procesos actuales que son: tratamiento de imágenes, conversión al HTML e Hipervínculo y publicación. Para ambos se toma como requisito el libro en formato Microsoft Word.

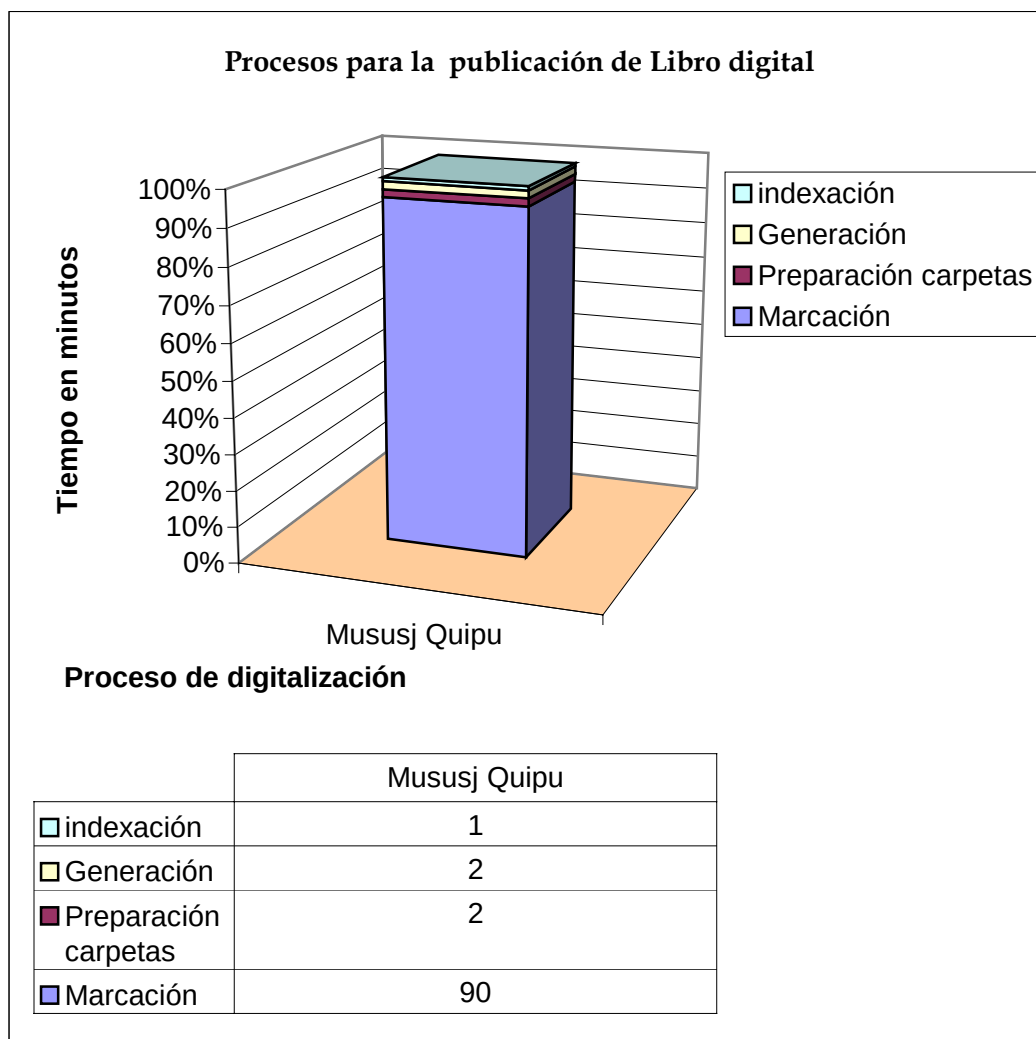


Figura 24: Procesos de sistema Musuj Quipu y sus tiempos

2.- Problemas encontrados a la hora de realizar el marcado de libros:

No todos los tipos de libros se han podido generar por que tienen especificaciones propias, y no cumplen con un solo tipo de estándar , por ejemplo se ha tenido problemas en la generación de un libro de cuentos , que incluía conversaciones de personajes en prosa, las marcas para este tipo de libros no están definidas en la plantilla.

En el documento Word se asignan las marcas correctamente pero cuando se abre con el OpenOffice, estos varían en algunos aspectos, es recomendable borrar los errores antes de generar el libro, para evitar fallas.

Se debe marcar bien los libros, todo el contenido debe estar bien marcado con el tipo e letra del “bodytexto” y que no sea de ningún comprimido de texto porque generará con errores.

8.1.3. Análisis de tiempo de búsqueda

Se realizó este tipo de análisis para ver que tan eficiente fue la alternativa de usar los sistemas de recuperación de datos, si en realidad se aprovechan las ventajas proporcionadas, el tiempo de recuperación de información se muestra en la misma página por cada consulta, el tiempo de búsqueda varía dependiendo la cantidad de documentos indexados que tiene. Para 15 libros indexados se ha buscado la palabra “Inca” y se demoró 12.578125 segundos tal como se observa en la [Figura 25].

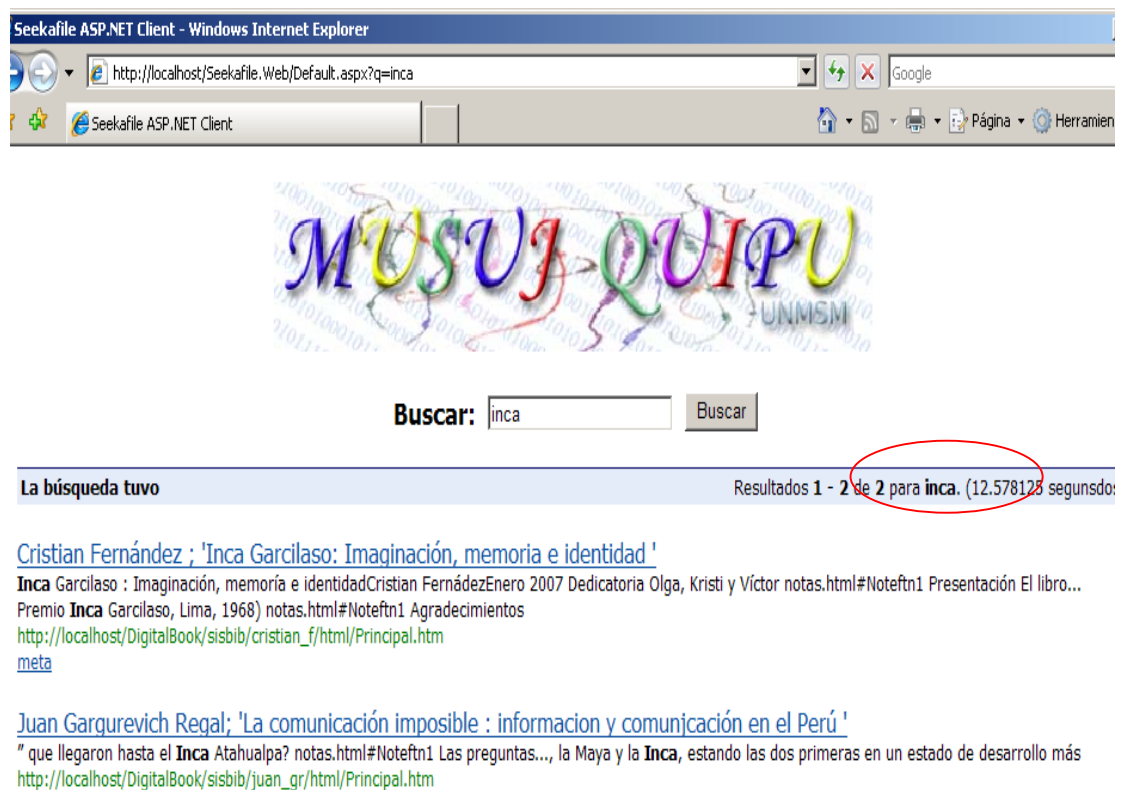


Figura 25: Tiempo de demora del buscador en encontrar resultado.

8.2. Trabajos futuros

- De acuerdo a las pruebas se ha visto que aun no se han creado las marcas para todo tipo de libros y un trabajo futuro seria agregar las marcas para todos los tipos de libros.
- Se debería incluir todo de tipo Material bibliográfico no solo libros si no también revistas y periódicos que tenemos en la hemeroteca.
- Se debe adoptar la metodología Musuj Quipu para la implementación de todo tipo de recurso como sonido, imágenes, gráficos, etc.
- Mejorar los algoritmos del buscador propuesto para que sea más robusta y personalizado.
- Se debe continuar con el mejoramiento del sistema hasta cubrir con la publicación de todos los tipos de materiales y así evolucionar la biblioteca digital Pedro Zulen a un ambiente hipertextual.

Capítulo IX: Conclusiones

A lo largo de este trabajo de tesis se ha visto el planteamiento de la metodología Musuj Quipu y su implementación, como puntos importantes se recalca en lo siguiente:

La metodología Musuj Quipu permite la preservación, publicación digital de ediciones de libros completos, organización en base de datos del metadato y organización de archivos de textos completos, indexación y recuperación eficiente e inmediata de información por contenido.

El uso de estándares internacionales nos permite preservar, simplificar la interoperabilidad y facilita el acceso de diferentes sistemas sin importar la ubicación física. Por ello en este trabajo de tesis se usó los estándares TEI, DC, XML y HTML.

Los metadatos tienen gran importancia en la composición de las bibliotecas digitales y es una buena práctica la creación del metadato en el momento mismo de producción de los documentos.

El mecanismo usado para la recuperación de información de los libros almacenados, se basa en el indexamiento de términos y el ordenamiento por relevancia de los resultados según el modelo vectorial.

Para el desarrollo de sistema Musuj Quipu, se diseñó una arquitectura de varias capas como: capa de datos, capa de negocios y capa de interfaz de usuario.

El desarrollo del aplicativo “Sistema Generador” es construido con el objetivo de procesar el documento marcado y obtenerlo en el formato HTML e ingresar el metadato del libro.

Construcción del interfaz del sistema Musuj Quipu en web usando ASP.NET, este sistema permite al usuario realizar la búsqueda del acervo bibliográfico que fue procesado mediante el “Sistema Generador” y se ha indexado con el componente que se ejecuta como Servicio Windows.

Las pruebas realizadas sirvieron para medir la eficiencia del sistema “Musuj Quipu” en cuanto a su robustez, independencia, integridad, seguridad y velocidad de recuperación. Demostrando la que la metodología propuesta es la adecuada para este tipo de sistemas. Su aplicación, implementación y seguimiento es de fácil uso para el usuario.

Para poner en marcha una biblioteca digital se requiere la participación y dedicación de diversos tipos de profesionales trabajando en equipo, también se necesita muchos recursos tecnológicos y humanos. Es necesario poner un cuidado especial en la definición de los aspectos técnicos como la colección, software y hardware.

Es necesario un buen análisis de la estructura de datos que va soportar una biblioteca virtual, y que debe permitir una óptima recuperación.

Es importante que la información este a disposición de todos para su uso en las diversas investigaciones.

Glosario

API: (Application Programming Interface - Interfaz de Programación de Aplicaciones): “Es un conjunto de especificaciones de comunicación entre componentes software. Se trata del conjunto de llamadas al sistema que ofrecen acceso a los servicios del sistema desde los procesos y representa un método para conseguir abstracción en la programación, generalmente (aunque no necesariamente) entre los niveles o capas inferiores y los superiores del software. Uno de los principales propósitos de una API consiste en proporcionar un conjunto de funciones de uso general, por ejemplo, para dibujar ventanas o iconos en la pantalla. De esta forma, los programadores se benefician de las ventajas de la API haciendo uso de su funcionalidad, evitándose el trabajo de programar todo desde el principio.”[56]

ASCII (American Standard Code for Information Interchange): La norma internacional de facto para los números codificados utilizados por los ordenadores para representar las letras mayúsculas y minúsculas del alfabeto latino, los números, puntuación, etc.

Base de Datos: Una base o banco de datos es un conjunto de datos que pertenecen al mismo contexto almacenados sistemáticamente para su posterior uso. En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta.[58]

Bit (Binary DigIT): La unidad de medida de información equivalente a la elección entre dos posibilidades igualmente probables.

Byte: Unidad de información digital. 1 byte equivale a 8 bits. Un megabyte equivale a 1 millón de bytes.

Biblioteca digital: Es un repositorio de acervos y contenidos digitalizados, almacenados en diferentes formatos electrónicos por lo que el original en papel, en caso de existir, pierde supremacía. Generalmente, son bibliotecas pequeñas y especializadas, con colecciones limitadas a sólo algunos temas.

Captura de la imagen: Proceso por el que se obtiene una representación digital de un original constituida por un conjunto de elementos pictóricos o "píxels" mediante el escaneado o fotografía digital.

CSIC : Consejo Superior de Investigaciones Científicas.

Digitalización: Proceso por el cual el contenido físico de un material pasa a ser digital.

DC (dublin core) : Estándar de metadatos.

Documento: Item de almacenamiento y recuperación.

Documento estructurado: Es un documento con una división (Autor, fecha de creación, lugar,...) o describiendo la jerarquía de contenidos (capítulos, secciones, subsecciones,...).

Documento de texto completo: Documento almacenado en un SARI con el contenido completo sin resumir.

DSSSL: DSSSL (Document Style Semantics and Specification Language) es un lenguaje para describir estilos en documentos. Se utiliza junto con jade/openjade para transformar documentos DocBook SGML y XML en archivos pdf o html. Actualmente los documentos SGML y XML se pueden transformar a otros formatos como ps, rtf, doc, etc.

Hipertextos: Un hipertexto es un documento digital o no, que se puede leer de manera no secuencial. Un hipertexto tiene los siguientes elementos: secciones, enlaces o hipervínculos y anclajes. Las secciones o nodos son los componentes del hipertexto o hiperdocumento. Los enlaces son las uniones entre nodos que facilitan la lectura secuencial o no secuencial del documento. Los anclajes son los puntos de activación de los enlaces.

HTML HyperText Markup Language: Lenguaje de codificación que se utiliza para crear documentos hipertexto de uso en la World Wide Web
<http://www.w3.org/MarkUp>

Información: Es un conjunto ordenado de datos los cuales son manejados según la necesidad del usuario, para que un conjunto de datos pueda ser procesado eficientemente y pueda dar lugar a información, primero se debe guardar lógicamente en archivos.

JPEG Joint Photographic Experts Group: Norma para la compresión de la imagen con calidad fotográfica en la World Wide Web. El formato JPEG se prefiere al formato GIF para imágenes fotográficas como opuesto al sistema de líneas.<http://www.jpeg.org/>

Libro Digital: Es el contenido de todo un libro pero en formato digital, respetando su orden físico. Ebokks.

Lenguaje de marcado: Lenguaje utilizado para especificar fuentes del documento alojado en la World Wide Web mediante etiquetas embebidas en el documento.

MARC (Machine Readable Cataloguing): Los formatos MARC son normas para la representación y comunicación de información bibliográfica y relacionada de forma legible por ordenador.

<http://lcweb.loc.gov/marc/umb/um01to06.html>

Metadato: Datos sobre datos. Por lo general, incluye información sobre el contenido intelectual de la imagen, sobre la representación de los datos digitales e información sobre la gestión de derechos o de la seguridad.

Music-Brainz : Música. Basado en XML/RDF.

Musuj Quipu : Frase quechua, en traducción separada (Musuj : Nuevo, Quipu : Nudos, Para los Incas sistema para llevar la cuenta de sus estadísticas y guardar su información).

Significado asignado en este trabajo de tesis es nueva forma de almacenamiento para la preservación digital.

NAMESPACES: Espacios de nombres, Permite que el autor de un documento XML le diga al analizador o parser qué DTD usar cuando analice un elemento, este DTD o Esquema puede estar en su propio documento, externo o en en la web.

NBII: Para información biológica. Complejo. Administración de la clasificación taxonómica.

ISO 8601: “La norma ISO 8601 especifica la notificación estándar utilizada para almacenar las fechas. Esta notación facilita la migración entre distintas plataformas. El estándar internacional para la notación de la fecha es: AAAA-MM-DD” [61]. Por ejemplo 23 de mayo de 2007 se escribe 2007-05-23.

OAIS : Open Archival Information System. Véase el borrador del Modelo de Referencia OAIS en <http://www.ccsds.org/documents/pdf/CCSDS-650.0R-2.pdf>.

Objetos: Entidad informativa que puede ser manipulada individualmente. Cabe que sea información “primaria” de cualquier tipo o información Sobre otra información (metadatos).

Palabra clave: Una palabra elegida por el autor, editor o automáticamente, para representar el contenido del documento.
Sinónimo: término.

Parser: “Programa o módulo encargado del parseo de un texto” [60].

Parseo: “Parcing. Proceso de analizar una secuencia de símbolos a fin de determinar su estructura gramatical con respecto a una gramática formal dada. Formalmente es llamado análisis de sintaxis. Un parseador (parser) es un programa de computación que lleva a cabo esta tarea. El parseo transforma una entrada de texto en una estructura de datos (usualmente un árbol) que es apropiada para ser procesada. Generalmente los parseadores primero identifican los símbolos de la entrada y luego construyen el árbol de parseo para esos símbolos” [60].

PICS (Platform for Internet Content Selection): Mecanismo que utiliza metadatos para controlar el acceso a determinado tipo de páginas atendiendo a un sistema de clasificación previamente establecido.

P3P (Plataforma para las Preferencias de Privacidad): Es un lenguaje estándar que ofrece a los usuarios una forma sencilla y automatizada de controlar en mayor medida el uso que se hace de su información personal en los sitios Web que visitan.

Query String: Cadena de consulta, este término generalmente se utiliza para hacer referencias a una interacción con una base de datos.[58]

SARI (Sistema de Almacenamiento y recuperación de Información) es información no estructurada. Hay más información no estructurada (libros, artículos, revistas, notas, e-mails, etc.) que información estructurada.

SISBIB: Sistema de bibliotecas de la universidad Nacional Mayor de San Marcos.

SGBD (Sistema de gestión de bases de datos) tiene un lenguaje de descripción de datos no ambiguo. Se suele expresar que la información de un SGBD es información estructurada, tiene un lenguaje de consulta formal no ambiguo.

SOIF (Summary ObjectInterchange Format - en Harvest) e **IAFA** (Internet Anónimo FTP Archive).- como ficheros de intercambio en sistemas de indización distribuida

W3C: El [Consorcio World Wide Web \(W3C\)](#) desarrolla tecnologías inter-operativas (especificaciones, líneas maestras, software y herramientas) para guiar la Red a su potencialidad máxima a modo de foro de información, comercio, comunicación y conocimiento colectivo.

Referencias bibliográficas

1. Gladney, Fox, Ahmed, Ashany, Belkin y Zemankova. Digital Library. Gross structure and requirements. Report from a March 1994 Workshop Proceeding of Digital Librerries, pág. 101-107.
2. Negroponte Nicholas, El mundo digital (1995) , ediciones B(2000), 288p.
3. Ortiz-Repiso, V.2 (1999). "Nuevas perspectivas para la catalogación: metadatos versus MARC". Revista española de Documentación.
4. Chung-Sheng Li. Stone, H.S. "Digital Library Using Next Generation Internet". IEEE Communications Magazine 37(1) (January 1999).
Dowling, Thomas. "Using Multimedia File Formats".
5. Fink, Donald G. "The Impact of Technology on Library Science". Special Libraries. 1977, February.
6. Walt Crawford y Michel Gorman, Future Libraries: dreams, madness & reality. Chicago: American Library Association, 1995.
7. Hecktart, Ronald J. "Machine Help and Human Help in the Emerging Digital Library. College and Research Libraries. 1998, May, page. 250-259.
8. Ricardo Bayes, Berthier Ribeiro-Neto. "Modern Information Retrieval".1999.
9. Lowry, Charles B. "Preparing for the Technologic Future: A Journey of Discovery. A: Library Hi Tech. 1995. Pág. 39-53

10. Mintzer, Fred. "Developing Digital Libraries of Cultural Content for Internet Access". A: IEEE Communications Magazine. 1999. Vol. 37, Núm. 1, Pág. 72-78
11. Mitchell, William J. "Homer to Homer-Page: Designing Digital Books. 1996, February.
12. Chung-Sheng Li. Stone, H.S. "Digital Libraries Using Next Generation Internet". A: IEEE Communications Magazine. 1999. Vol. 37, Núm. 1, pàg. 70-71
13. Watstein, S.; Calarco, P. Y Graphery, J.S (1999). "Digital Library: keywords". Reference Services Review, 27(4), p.344-352
14. Future libraries: beyond the walls (1995).p. 165-177. En: W. Cranford y M. Gorman (eds.).Chicago: ALA.
15. Garduño, V.R. (1998). "Organización de la información documental y su utilidad social". t. T, p.40-64. En La información en el inicio de la era electrónica México: CUIB, UNAM.
16. Pontificia Universidad Javieriana: Facultad de comunicación y lenguaje "Revista interamerican de Nuevas Tecnologías de la información". Vol.7. nº3.
17. Fox, Edward. Kipp, Neill. MATHER, Paul. "How Digital Libraries Will Save Civilization". A: Database Programming & Design, 1998 Vol. 11. Núm. 8. Pág. 61-65. Disponible en : <http://www.dbpd.com>
18. Garret, John. "Digital Libraries: The Grand Challenges". EDUCOM Review, 1993, Vol. 28, number 4. Disponible en: <http://www.ifla.org/documents/libraries/net/garrett/txt>
19. Graham, Peter. "Requirements for the Digital Research Library". College and Research Libraries, July 1995, pàg.331-339. Disponible en : <http://www.ifla.org/documents/libraries/net/drc.htm>

20. Litman, Jessica. "Copyright Law and Electronic Access to Information". A: Peer-Reviewed Journal on the Internet. Disponible en: <http://www.ifl.org>
21. Label France, la revista, André Maxlaur (09/1996) .nº25. Disponible en:
http://www.diplomatie.gouv.fr/label_france/ESPANOL/LETTR ES/MALRAUX/malraux.html. [Consultado en diciembre del 2005].
22. Colegio San Martín de Tours. Disponible en:
<http://www.smt.edu.ar/>. [Consultad en diciembre 2005].
23. Comisión de Preservación y Acceso, Washington- DC 20036-2217 (202) 939-3400 .Disponible en:
<http://www.clir.org/pubs/reports/dpcmiss/dpcmiss.html>. [Consultado en diciembre 2005].
24. Ley de democratización del libro y de fomento de la lectura. Disponible en:
http://www.cerlalc.org/documentos/peru_libro.htm. [Consultado en diciembre 2005].
25. Kuny, Terry, Cleveland, Gary (2001). La biblioteca digital: Mitos y retos. Métodos de Información, #48 -noviembre 2001, p. 72-81. Disponible en: <http://www.metodosdeinformacion.com>. [Consultado en diciembre 2005].
26. Bunker, Geri. ZICK, Greg. "Collaboration as a key to Digital Library Development". A: D-Lib Magazine. 1999, March. Disponible en: <http://www.dlib.org>
27. Morales Salcedo R. Bibliotecas Digitales, disponible en:
http://www.pue.udlap.mx/~tesis/msp/morales_s_r/capitulo2.pdf. [Consultado: 20 de mayo 2005].
28. Antonio M. Battro & Percival J. DENMA, La educación Digital, capítulos 2, 3, 5, 6. disponible en:

- <http://www.byd.com.ar/ed8www2.htm>. [Consultado en diciembre 2005].
29. Association of Research Libraries, Biblioteca Digital, disponible en: <http://www.arl.org/index.html>. [consultado en enero 2005].
30. York V. A biblioteca digital. Recursos y proyectos. Disponible en: <http://www.cg.org.br/gt/gtbv/YorkatIBICT/tsld004.htm> [Consultado: 15 de abril del 2005].
31. Dr. Carlos castillo, Tecnologías de información, procesamiento de texto XML. Disponible en :
http://www.tejedoresdelweb.com/307/articles-59683_Modelo_Datos_XML.pdf. [Consultado: marzo 2007].
32. Introducción a XML ,disponible en :
<http://geneura.ugr.es/~jmerelo/xml/>. [Consultado en marzo 2007].
33. Arbis Percy Reyes Paredes, Arquitectura XML .net framework, disponible en:
http://www.elguille.info/colabora/NET2005/Percynet_ArquitecturaXML_en_NETFramework.htm : [consultado en octubre 2005.]
34. Miguel Angel Alvarez, Tutorial de XML, Disponible en:<http://www.desarrolloweb.com/articulos/450.php?manual=18..> [consultado en noviembre 2005].
35. Brais Méndez Ferreiro, XML, Disponible en:
<http://www.elrincondelprogramador.com/default.asp?pag=articulos/leer.asp&id=15>. [Consultado en abril 2005].
36. María Isabel García Arenas, Curso XLM 1ra edición. Disponible en:
<http://geneura.ugr.es/~maribel/xml/introduccion/index.shtml#11>. [consultado en julio del 2005].
37. W3C, Espacio de nombres en XML. Disponible en:
<http://html.conclase.net/w3c/xml-names-es/#sec-intro> . [Consultado en julio del 2005].

38. Alfredo Reino, Tutorial de XML. Disponible en:
<http://www.asptutor.com/xml/index1.asp#atributos>.
[Consultado en mayo del 2005].
39. *The World Wide Web Consortium (W3C)*, Disponible en:
<http://www.w3.org/> . Los creadores de los estándares HTML, XML, XSL, XHTML y muchos otros; [Consultado en abril de 1 2005].
40. Diane Hillmann, Usando Dublin Core, Disponible en:
<http://es.dublincore.org/documents/2001/04/12/usageguide/> .
[Consultado en agosto del 2005].
41. José A. Senso, Antonio de la Rosa Piñero, El concepto de metadato. Algo más que descripción de recursos electrónicos. Disponible en: <http://www.scielo.br/pdf/ci/v32n2/17038.pdf> .
[Consultado en Junio del 2005].
42. Digital Libraries Initiative Phase 2. Disponible en:
<http://www.nsf.gov/pubs1998/nsf9863/nsf9863.htm>
43. Digital Library Project. University of California, Berkeley.
Disponible en: <http://elib.cs.berkeley.edu>
44. Victor Fresno Fernández, Tecnologías relacionadas con el XML,
Disponible en :
http://frodo.escet.urjc.es/~oa/material/TecnologiasXML_OA_04-05.pdf. [Consultado en marzo 2007].
45. Howe, D. *Free on-line Dictionary of Computing (FOLDOC)*. 1993.
Disponible en: <http://wombat.doc.ic.ac.uk/foldoc/> . [Consultado en noviembre del 2005].
46. Ora Lassila, Ralph R. Swick, Resource Description Framework (RDF), Disponible en:
<http://www.sidar.org/recur/desdi/traduc/es/rdf/rdfesp.htm>.
[Consultado en setiembre del 2005].

47. Javier Solá, "Tejiendo la Red, Tim Berners - Lee". Disponible en:
http://www.sigloxxieditores.com/Libros/tejiendo_la_red.htm.
[Consultado en diciembre 2005].
48. RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax,
Disponible en: <http://www.faqs.org/rfcs/rfc2396.html>.
[consultado en noviembre del 2005].
49. Referencias en HTML, Stefan Muñiz (2001) Traducción: Virgilio
Krumbacher (2002), Disponible en:
<http://es.selfhtml.org/html/allgemein/referenzieren.htm#uri> .
[Consultado en octubre del 2005].
50. Joaquin Bravo Montero, "Xslt, XSL, Xpath: ¿Qué es cada cosa?",
Disponible en: <http://www.programacion.net/articulo/xslt/>,
[Consultado en noviembre 2005].
51. <http://www.sidar.org/recur/desdi/traduc/es/rdf/rdfesp.htm>
National Digital Library Federation. "America's Heritage: Mission
and Goals for a National Library Federation."
<http://www.ifla.org/documents/libraries/net/digfed2.htm> .
[Consultado en diciembre 2005].
52. NEWBY, Gregory B. "The Digital Electronic Library". Syracuse
University: School of Information Studies, 1991. Disponilbe en:
<http://www.ifla.org/documents/libraries/net/newby.txt> . [
Consultadoe ne mayo 2006]
53. Dora Pérez, Biblioteca de la Universitat Oberta de Catalunya, "La
biblioteca digital" disponible en:
http://www.nuestraldea.com/reflexiones/biblioteca_digital.htm
[Consultado en agosto del 2006]
54. Berta Araujo, "Motor de búsqueda para un SRI con
agrupamiento", apéndice del Tomo I, disponible en :
<http://trevinca.ei.uvigo.es/~pcuesta/sm/practicas/Lucene.pdf> [en PDF].

55. "Wikipedia" , la enciclopedia libre , disponible en:
<http://es.wikipedia.org/wiki/Lucene> [consultado en noviembre 2006]
56. Recuperaación de información, disponible en:
<http://ict.udlap.mx/people/carlos/is215/ir12.html> y
<http://ict.udlap.mx/people/carlos/is346/admon08.html>
[consultado en noviembre 2007].
57. Wikipedia, enciclopedia libre, disponible en:
http://es.wikipedia.org/wiki/Query_string [consultado en Marzo 2007].
58. IBM. Delve inside the Lucene indexing mechanism, disponible en:
<http://www-128.ibm.com/developerworks/library/wa-lucene/>
[consultado en marzo del 2007].
59. IBM. Lucene, disponible en:
<http://www.ibm.com/developerworks/cn/java/j-lo-lucene1/>
[consultado en marzo 2007].
60. Diccionario Informático, disponible en:
<http://www.alegsa.com.ar/Dic/parseo.php> [Consultado en marzo del 2007].
61. Norma ISO 8601, disponible en: <http://www.f-integra.org/plataforma/iso8601.htm> [Consultado en mayo del 2007].
62. José Abaitua, (versión reciclada de la "Revolución Web")
Funcionamiento TEI", Facultad de Filosofía y Letras, [Universidad de Deusto](http://www.uned.es), E-48080 Bilbao. Disponible en:
<http://paginaspersonales.deusto.es/abaitua/konzeptu/byte.htm#TEI> [Consultado en mayo del 2007].
63. Biblioteca de la Universidad de Cornell / Departamento de
Investigación, "Digitalización digital".Disponible en:
<http://www.library.cornell.edu/preservation/tutorial->

- [spanish/preservation/preservation-01.html](#) [Consultado en mayo 2007].
64. José Manuel Barruecos, Preservación y conservación de documentos digitales. Disponible en :
<http://www.archivovirtual.org/seminario/ediciencia/pdf/CAPITULO7.pdf> [Consultado en mayo del 2007].
65. National Library of Australia, Guidelines for the preservation of digital heritage.2003. Disponible en:
<http://unesdoc.unesco.org/images/0013/001300/130071e.pdf> [Consultado en mayo del 2007].
66. Clara Lopez Guzman, modelo para el desarrollo de Bibliotecas Digitales especializadas. Disponible en :
http://www.bibliodgsca.unam.mx/tesis/tes7cllg/sec_10.htm. [Consultado en mayo del 2007].
67. Grupo de Trabajo de Patrimonio Digital del Ministerio de Cultura y las Comunidades Autónomas, Jornadas en España sobre Preservación del Patrimonio Digital.Disponible en:
<http://www.abinia.org/boletin/5-1/miscelaneas-3.htm> [Consultado en junio 2007].
68. Joseba Abaitua, Universidad de Deusto, Perspectivas en el ámbito del patrimonio digital: Nuevo retos de futuro. Disponible en:
http://www.kultura.ejgv.euskadi.net/r46-6614/eu/contenidos/informacion/conferencia_politica_culturale/eu_10312/adjuntos/ponencia_abaitua_cas.pdf [Consultado en junio 2007].
69. José Manuel Barruecos, Preservación y conservación de documentos digitales. Disponible en :
<http://www.archivovirtual.org/seminario/ediciencia/pdf/CAPITULO7.pdf> [Consultado en junio 2007].

70. Preservación Digital, Universidad de Cornell. Disponible en:
<http://www.library.cornell.edu/preservation/tutorial-spanish/preservation/preservation-04.html> [Consultado en junio del 2007].
71. Lucen.NET, Apache. Disponible en:
<http://incubator.apache.org/lucene.net/> [Consultado en junio 2007].
72. La preservación digital y las universidades: el estado de la cuestión, Alice Keefer, Núrcia Gallard. Disponible en :
<http://eprints.rclis.org/archive/00004816/01/Fesabid2003preprint.pdf>
[Consultado en junio 2007].
73. Directrices para la preservación del patrimonio digital – Proyecto de carta de la UNESCO para la preservación del patrimonio Digital, Biblioteca nacional de Australia. Capítulo 2 Pag. 11. Disponible en :
<http://unesdoc.unesco.org/images/0013/001300/130071s.pdf>
[Consultado en Junio 2007].

Anexos

Anexo 1: Acuerdo de Cartagena

Legislación sobre derechos del autor de las obras publicadas en la web, en nuestro país se rige al acuerdo de Cartagena que dice lo siguiente:

Decisión 351: Régimen Común sobre Derecho de Autor y Derechos Conexos -Comisión del Acuerdo de Cartagena

17 de diciembre de 1993

RÉGIMEN COMÚN SOBRE DERECHO DE AUTOR Y DERECHOS CONEXOS

CAPÍTULO I: del Alcance de la Protección

Artículo 1.- Las disposiciones de la presente Decisión tienen por finalidad reconocer una adecuada y efectiva protección a los autores y demás titulares de derechos, sobre las obras del ingenio, en el campo literario, artístico o científico, cualquiera que sea el género o forma de expresión y sin importar el mérito literario o artístico ni su destino.

Asimismo, se protegen los Derechos Conexos a que hace referencia el Capítulo X de la presente Decisión.

Artículo 2.- Cada País Miembro concederá a los nacionales de otro país, una protección no menos favorable que la reconocida a sus propios nacionales en materia de Derecho de Autor y Derechos Conexos.

Artículo 3.- A los efectos de esta Decisión se entiende por:

Autor: Persona física que realiza la creación intelectual.

Artista intérprete o ejecutante: Persona que representa, canta, lee, recita, interpreta o ejecuta en cualquier forma una obra.

Autoridad Nacional Competente: Órgano designado al efecto, por la legislación nacional sobre la materia.

Copia o ejemplar: Soporte material que contiene la obra, como resultado de un acto de reproducción.

Derechohabiente: Persona natural o jurídica a quien por cualquier título se transmite derechos reconocidos en la presente Decisión.

Distribución al público: Puesta a disposición del público del original o copias de la obra mediante su venta, alquiler, préstamo o de cualquier otra forma.

Divulgación: Hacer accesible la obra al público por cualquier medio o procedimiento.

Emisión: Difusión a distancia de sonidos o de imágenes y sonidos para su recepción por el público.

Fijación: Incorporación de signos, sonidos o imágenes sobre una base material que permita su percepción, reproducción o comunicación.

Fonograma: Toda fijación exclusivamente sonora de los sonidos de una representación o ejecución o de otros sonidos. Las grabaciones gramofónicas y magnetofónicas se consideran copias de fonogramas.

Grabación Efímera: Fijación sonora o audiovisual de una representación o ejecución o de una emisión de radiodifusión, realizada por un período transitorio por un organismo de radiodifusión, utilizando sus propios medios, y empleada en sus propias emisiones de radiodifusión.

Obra: Toda creación intelectual original de naturaleza artística, científica o literaria, susceptible de ser divulgada o reproducida en cualquier forma.

Obra audiovisual: Toda creación expresada mediante una serie de imágenes asociadas, con o sin sonorización incorporada, que esté destinada esencialmente a ser mostrada a través de aparatos de proyección o cualquier otro medio de comunicación de la imagen y de

sonido, independientemente de las características del soporte material que la contiene.

Obra de arte aplicado: Creación artística con funciones utilitarias o incorporada en un artículo útil, ya sea una obra de artesanía o producida en escala industrial.

Obra Plástica o de bellas artes: Creación artística cuya finalidad apela al sentido estético de la persona que la contempla, como las pinturas, dibujos, grabados y litografías. No quedan comprendidas en la definición, a los efectos de la presente Decisión, las fotografías, las obras arquitectónicas y las audiovisuales.

Oficina Nacional Competente: Órgano administrativo encargado de la protección y aplicación del Derecho de Autor y Derechos Conexos.

Organismo de radiodifusión: Empresa de radio o televisión que transmite programas al público.

Productor: Persona natural o jurídica que tiene la iniciativa, la coordinación y la responsabilidad en la producción de la obra, por ejemplo, de la obra audiovisual o del programa de ordenador.

Productor de fonogramas: Persona natural o jurídica bajo cuya iniciativa, responsabilidad y coordinación, se fijan por primera vez los sonidos de una ejecución u otros sonidos.

Programa de ordenador (Software): Expresión de un conjunto de instrucciones mediante palabras, códigos, planes o en cualquier otra forma que, al ser incorporadas en un dispositivo de lectura automatizada, es capaz de hacer que un ordenador –un aparato electrónico o similar capaz de elaborar informaciones–, ejecute determinada tarea u obtenga determinado resultado. El programa de ordenador comprende también la documentación técnica y los manuales de uso.

Publicación: Producción de ejemplares puestos al alcance del público con el consentimiento del titular del respectivo derecho, siempre que la

disponibilidad de tales ejemplares permita satisfacer las necesidades razonables del público, teniendo en cuenta la naturaleza de la obra.

Retransmisión: Reemisión de una señal o de un programa recibido de otra fuente, efectuada por difusión inalámbrica de signos, sonidos o imágenes, o mediante hilo, cable, fibra óptica u otro procedimiento análogo.

Titularidad: Calidad del titular de derechos reconocidos por la presente Decisión.

Usos honrados: Los que no interfieren con la explotación normal de la obra ni causan un perjuicio irrazonable a los intereses legítimos del autor.

Uso personal: Reproducción u otra forma de utilización, de la obra de otra persona, en un solo ejemplar, exclusivamente para el propio uso de un individuo, en casos tales como la investigación y el esparcimiento personal.

Dada en la ciudad de Lima, Perú, a los diecisiete días del mes de diciembre de mil novecientos noventa y tres.

Anexo 2: Resumen de modelos de recuperación de información [56]

En este anexo se describe los diversos modelos de recuperación de información para tener un mejor entendimiento de ellos.

2.1. Recuperación de Datos vs. Recuperación de Información

Recuperación de datos consiste en determinar que documentos contienen las llaves del query en el documento, no resuelve el problema de recuperar información acerca de un tema. Recuperación de información, es la representación y organización de la información, deben proveer al usuario un fácil acceso a su interés personal. Dado un query, la meta de una llave es recuperar la información relevante para el usuario.

2.2. Visión Lógica de los Documentos

Es la representación de documentos y/o páginas web que forman parte de una colección (Sistema de Recuperación de información [SRI]).

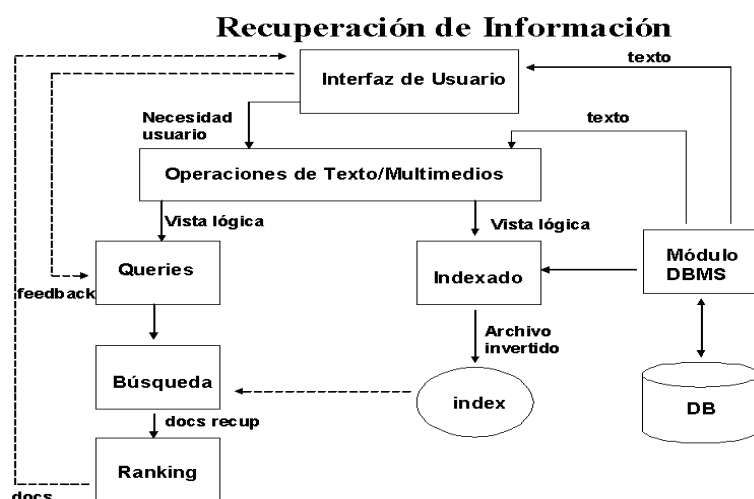


Figura 2.1: Arquitectura del SRI

La forma más común de representar un documento de texto es por un sistema de términos indexados o palabras llaves.

Estos términos son extraídos con el siguiente proceso:

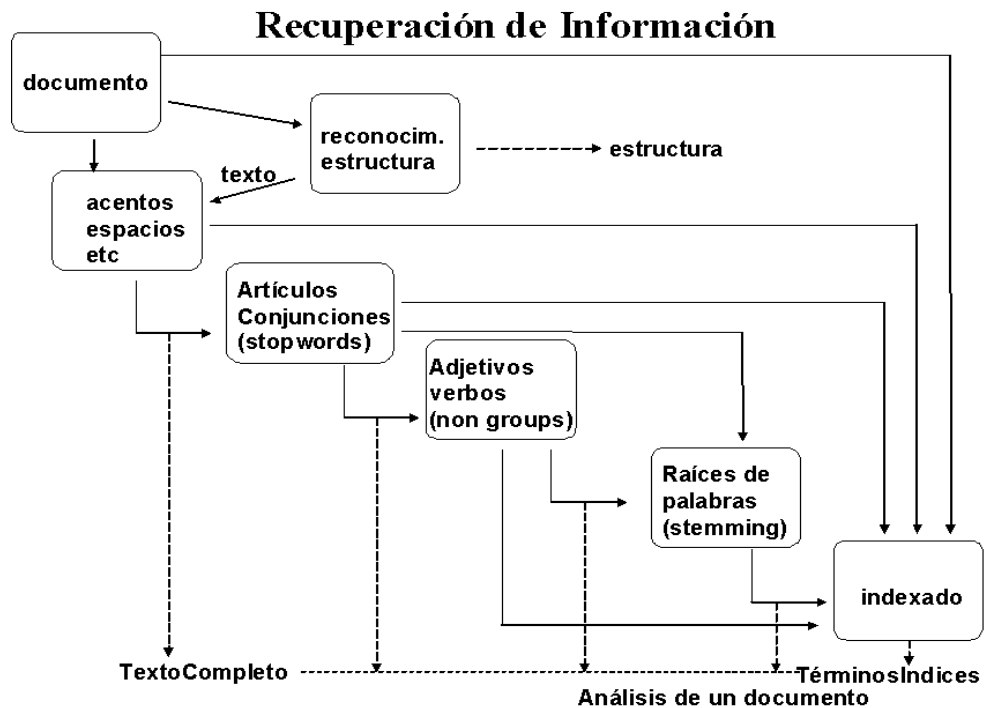


Figura 2.2: Procesamiento de la información en un SRI

Para la lematización los "stopwords" (palabras clave) generalmente se utilizan listas de palabras, ya sean de un lenguaje controlado (alguna especialidad) o bien de un idioma particular.

Por otro lado para realizar el "stemming" (reducir palabras de su raíz gramatical) se emplea el algoritmo de Porter, solo que hay que tener en cuenta el idioma a utilizar ya que por lo general se encuentra en inglés aunque hay versiones para español.

<http://www.tartarus.org/~martin/PorterStemmer/>

2.3. Tareas en la de recuperación de información

Existen 2 tipos de tareas principales en la recuperación de información:

- Las centradas en la computadora: consiste en construir un índice eficiente, procesar las consultas del usuario con un alto rendimiento y algoritmos (modelos de recuperación) que mejoren la calidad de respuesta del sistema.
- Las centradas en el humano: consiste principalmente en estudiar las necesidades del usuario, saber cómo afecta a la organización, operación y visualización de resultados en el sistema de recuperación.

2.4. Modelos de Recuperación

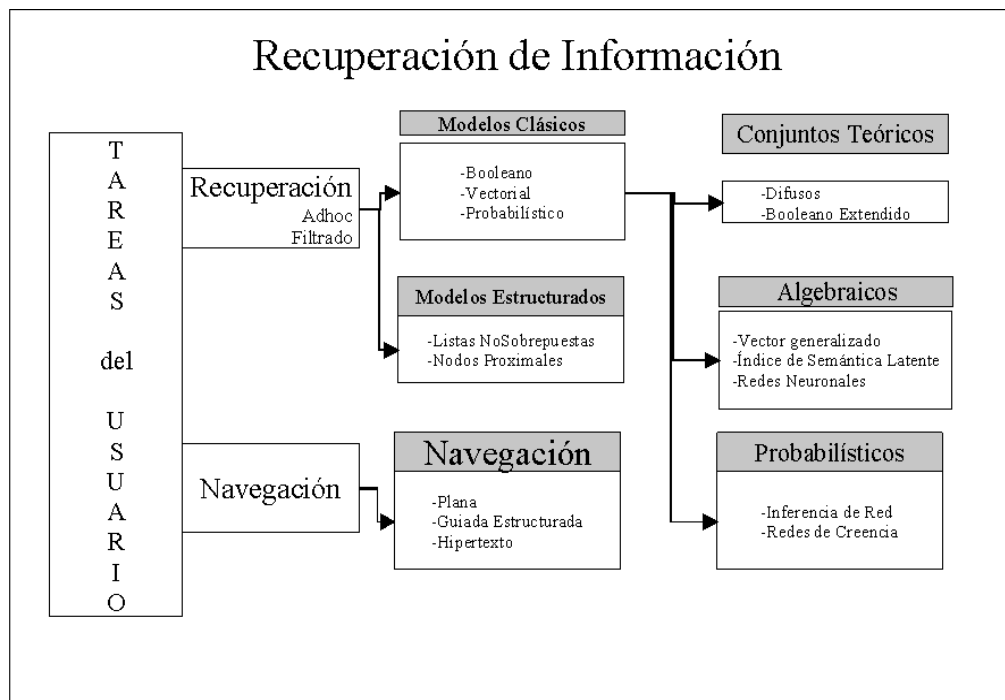


Figura 2.3: Clasificación de los sistemas de recuperación de información

2.4.1. Modelo Booleano

El modelo Boleano, es un modelo de recuperación simple basado en la teoría fija y álgebra de Boolean, este modelo proporciona un

grupo de trabajo que es fácil de usar por un usuario común de un sistema de IR. Además, las llamadas se especifican como expresiones de Boolean que tienen la semántica precisa.

Dado su simplicidad inherente y formalismo, el modelo de Boolean recibió la gran atención y se adoptó por muchos de los sistemas bibliográficos comerciales.

De este modelo se pueden destacar los siguientes puntos:

- La relevancia es binaria: un documento es relevante o no lo es.
- Consultas de una palabra: un documento es relevante si contiene la palabra.
- Consultas AND: Los documentos deben contener todas las palabras.
- Consultas OR: Los documentos deben contener alguna palabra.
- Consultas A BUTNOT B: Los documentos los documentos deben ser relevantes para A pero no para B.

Ejemplo: lo mejor de Maradona

Maradona AND Mundial AND ((México '86 OR Italia '12.)

BUTNOT U.S.A. '12.)

- Es el modelo más primitivo, sin embargo es el más popular.

¿Por qué es malo?

- No discrimina entre documentos más y menos relevantes.

- Da lo mismo que un documento contenga una o cien veces las palabras de consulta.
- Da lo mismo que cumpla una o todas las cláusulas de un OR.
- No permite ordenar los resultados.
- Puede resultar confuso

Ejemplo: Necesito investigar sobre los aztecas y los Incas

Aztecas AND Incas

(cuando en realidad lo que se pretendía era: Aztecas OR Incas).

¿Por qué es popular?

- Es una de los primeros modelos que se implementó y muchos de los primeros sistemas de IR se basaron en él
- La idea suele ser común entre los usuarios que la están usando.
- Es la opción favorita para insertar texto en un RDBMS (sistema administrador de bases de datos relacionales).
- Es simple de formalizar y eficiente de implementar.
- En algunos casos (usuarios expertos) puede ser adecuado.
- Puede ser útil en combinación con otro modelo ej. Para excluir documentos.
- Puede ser útil con buenas interfaces.

2.4.2 Modelo Probabilística

Este modelo fue introducido en 1976 por Roberston y Spark Jones y después se conoció como el modelo de la recuperación de independencia binario.

La idea fundamental es, dada una pregunta del usuario, se encuentra un conjunto de documentos que contienen los datos pertinentes, a este conjunto se le conoce como conjunto de la respuesta ideal.

El modelo sólo asume que esta probabilidad de relevancia depende de la pregunta y las representaciones del documento, que en este caso el usuario haga.

Ventajas:

- Se alinean los documentos en orden decreciente de su probabilidad de ser pertinentes (referenciados).

Desventajas:

- La necesidad de suponer la separación inicial de documentos en los conjuntos pertinentes y no pertinentes.
- No se toma en cuenta la frecuencia de un término del índice ocurre dentro de un documento (todo los pesos son binarios).
- Que adopta la independencia para las condiciones del índice.

Características

- Se presupone que existe exactamente un subconjunto de documentos que son relevantes para una consulta dada.
- Para cada documento, se intenta evaluar la probabilidad de que el usuario lo considere relevante.
- La relevancia de un documento se calcula como:

o $P(d \text{ relevante para } q) / P(d \text{ no relevante para } q)$

- Donde q es una pregunta del usuario y d los campos de cada documento.

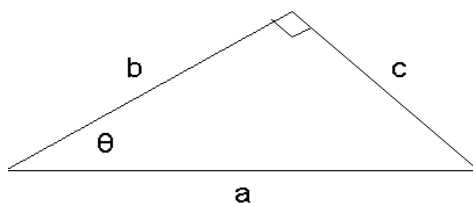
¿Por qué es poco popular?

- Se debe comenzar adivinando y luego refinar esa apuesta iterativamente.
- El modelo ve cada documento como un conjunto de términos.
- Necesita presuponer que los términos son independientes.
- Existen estudios que muestra que es inferior al modelo vectorial y casi todos los científicos lo consideran inferior.

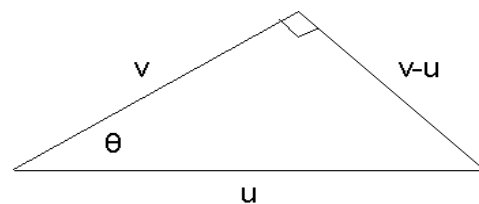
2.4.3 Modelo de Espacios Vectoriales

Es el modelo más popular hoy en día ya que permite discriminar correctamente entre documentos, veamos los conceptos iniciales:

Teoría de Vectores



$$c^2 = a^2 + b^2 - 2ab \cos \theta$$



$$u \cdot v = |u| |v| \cos \theta$$

Calcular el coseno del ángulo entre el vector $u=(1,2,3)$ y $v=(-4,-5,-6)$

$$\begin{aligned} |u| &= \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14} & |v| &= \sqrt{(-4)^2 + (-5)^2 + (-6)^2} = \sqrt{77} \\ u \cdot v &= 1(-4) + 2(-5) + 3(-6) = -32 \\ -32 &= \sqrt{14} \sqrt{77} \cos \theta = 7\sqrt{22} \cos \theta \\ \cos \theta &= -0.975 \end{aligned}$$

- Un documento d_i se modeliza como un vector

$$d_i \longrightarrow \vec{d}_i = (w(t_1, d_i), \dots, w(t_k, d_i))$$

donde $w(t_r, d_i)$ es el *peso* del término t_r en el documento d_i .

- Hay varias fórmulas para los pesos, una de las más populares es

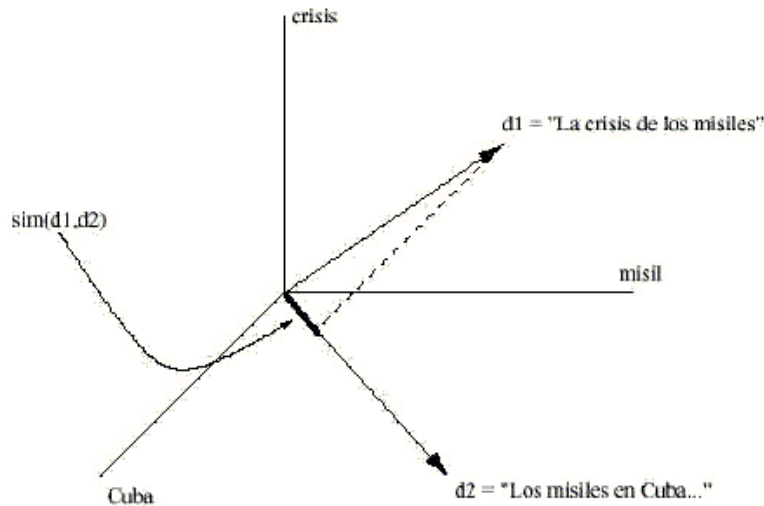
$$w(t_r, d_i) = w_{r,i} = \frac{tf_{r,i} \times idf_r}{|\vec{d}_i|} = \frac{tf_{r,i} \times \log \frac{N}{n_r}}{\sqrt{\sum_{s=1}^k (tf_{s,i} \times \log \frac{N}{n_s})^2}}$$

donde $tf_{r,i}$ (term frequency) es la cantidad de veces que t_r aparece en d_i , y n_r es la cantidad de documentos donde aparece t_r .

- Se calcula la *similitud* entre dos documentos mediante la *distancia coseno*:

$$sim(d_i, d_j) = \vec{d}_i \cdot \vec{d}_j = \sum_{r=1}^k w_{r,i} \times w_{r,j}$$

(que geométicamente corresponde al coseno del ángulo entre los dos vectores).



- Dado que en general cada palabra aparece una sola vez en la consulta y esta es muy corta, se puede en general calcular la relevancia de d_i para q con la fórmula

$$\text{sim}(d_i, q) = \sum_{r=1}^k w_{r,i} \times w_{r,q} \sim \sum_{t \in q} w_{t,i} \times \text{idf}_t$$

- La última fórmula no es idéntica a *sim* pero ordena los documentos de la misma forma (falta dividir por $|q|$).

Toma en cuenta:

- tf (term frequency): la frecuencia de un término en un cada documento
- maxtf (max term frequency): de un término entre todos los documentos
- f (frequency): es la normalización de la frecuencia, tf/maxtf
- $\text{idf} = \log(N/n_i)$ Frecuencia inversa, N -número de documentos en la colección, n_i -número de documentos donde aparece un término
- $w = f \times \text{idf}$, el peso de un término en un documento

Proceso

- Se selecciona un conjunto de palabras útiles para discriminar (términos o keywords).
- Se puede enriquecer esto con un proceso de lematización (o stemming), etiquetado, e identificación de frases.
- En los sistemas modernos, toda palabra del texto es un término, excepto posiblemente las stopwords o palabras vacías.
- Si un término aparece mucho en un documento, se supone que es importante en ese documento (tf crece).

- Pero si aparece un muchos documentos, entonces no es útil para distinguir ningún documento de los otros (idf decrece).
- Además normalizamos los módulos de los vectores para no favorecer documentos más largos.
- Lo que se intenta medir es cuánto ayuda ese término a distinguir ese documento de los demás.
- La similaridad es un valor entre cero y uno.
- Notar que dos documentos iguales tienen similaridad 1, y ortogonal (si no comparten términos) tienen similaridad cero.
- En particular, una consulta se puede ver como un documento (formado por esas palabras) y por lo tanto como un vector

El modelo es más general, y permite cosas como:

- Que la consulta sea un documento.
- Hacer clustering de documentos similares.
- Relevance feedback ("more like this").

2.5. Modelos basados en estructuras de texto

Introducción

- Problema:
Encontrar los documentos que contengan la cadena "holocausto atómico" con letra cursiva, y que se encuentre cerca de una figura cuya etiqueta dice "tierra".
- Solución:
Un modelo que permita la siguiente consulta:
misma-pagina(cerca_de("holocausto atómico", figura(etiqueta("tierra"))))

Definición:

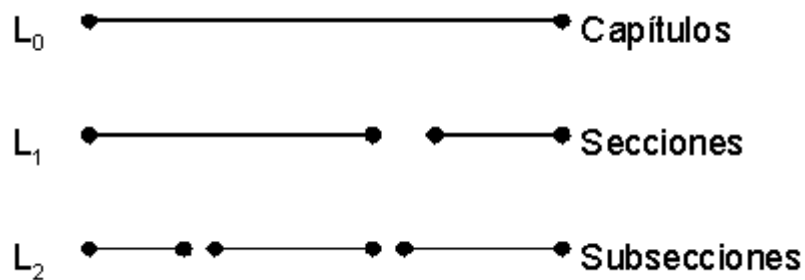
“Un modelo de RI que combina la información del contenido del texto con la información sobre la estructura del documento”.

Desventaja: no tiene una manera de clasificar los resultados en base a su importancia (ranking).

2.5.1. Modelos basados en listas no sobrepuestas (non-overlapping)

Idea:

Dividir el texto de cada documento en regiones que no están sobrepuestas y juntarlos en una lista.

**Implementación:**

Se crea un archivo invertido en el que cada componente estructural es una entrada en el índice. Asociado con cada una de estas entradas, hay una lista de regiones de texto como una lista de ocurrencias.

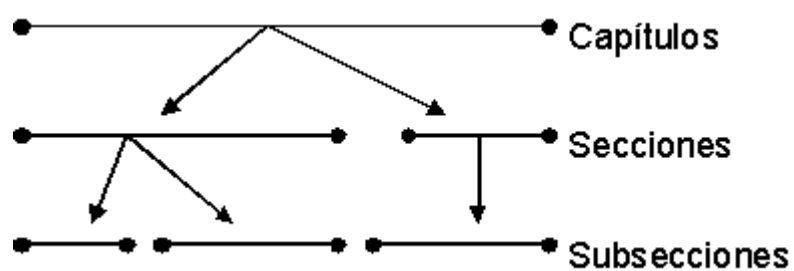
Ejemplos de consultas:

- seleccionar una región que contenga una palabra dada
- seleccionar la región A que no contenga una región B.

2.5.2 Modelos basados en nodos proximales

Idea:

Definir estructuras de indexamiento jerárquicas e independientes sobre un mismo documento.

**Implementación:**

Primero buscar los componentes que coinciden con la cadena especificada en la consulta y, subsecuentemente, evaluando cuál de estos componentes satisface la parte estructural de la consulta.

Ejemplos de consultas: (*section) with ('holocausto') .

Anexo 3: Hojas de estilo XSL creados para la implementación del sistema.

En este anexo se muestra los principales archivos XSLs creados para la generación de documentos HTML a partir del XML estructurado con el estándar TEI.

3.1. XSL para la cabecera de la página que corresponde a cada libro.

```
<?xml version="1.0" encoding="ISO-8859-15" ?>
<xsl:stylesheet
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="1.0">
  <xsl:output method="html" encoding="ISO-8859-15" />
  <xsl:template match="/">
    <html>
      <head>
        <title>Cabecera</title>
      </head>
      <body background="imagenes/barra.jpg">
        <xsl:apply-templates select="//principal" />
      </body>
    </html>
  </xsl:template>
  <xsl:template match="//principal">
    <xsl:element name="table">
      <xsl:attribute name="border">0</xsl:attribute>
      <xsl:attribute name="width">100%</xsl:attribute>
      <xsl:attribute name="align">left</xsl:attribute>
      <xsl:element name="tr">
        <xsl:element name="td">
          <xsl:attribute
            name="height">20%</xsl:attribute>
          <xsl:attribute
            name="width">20%</xsl:attribute>
          <xsl:element name="image">
            <xsl:attribute name="src">
              <xsl:text>imagenes/cabecera.
                gif</xsl:text>
            </xsl:attribute>
            <xsl:attribute
              name="align">baseline</xsl:attribute>
          </xsl:element>
          <xsl:element name="td" />
        </xsl:element>
      </tr>
    </table>
  </xsl:template>
</xsl:stylesheet>
```

```

</xsl:element>
= <xsl:element name="td">
  = <font name="verdana" color="#000000"
    size="3" style="bold" face="Arial">
    <xsl:attribute
      name="width">240</xsl:attribute>
    = <b>
      = <i>
        <xsl:value-of
          select="titulo / text()" />
        </i>
      </b>
    </font>
    <p />
  </xsl:element>
</xsl:element>
</xsl:element>
</xsl:template>
</xsl:stylesheet>

```

3.2. XSL para la bibliografía

```

<?xml version="1.0" encoding="ISO-8859-15" ?>
<xsl:stylesheet
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="1.0">
  <xsl:output method="html" encoding="ISO-8859-15" />
  <xsl:template match="/">
    <html>
      <head>
        <title>Bibliografia</title>
      </head>
      <body>
        <xsl:apply-templates select="//back/div/head" />
        <xsl:apply-templates select="//back/div/bibl" />
      </body>
    </html>
  </xsl:template>
  <xsl:template match="//back/div/head">
    <xsl:element name="table">
      <xsl:attribute name="border">0</xsl:attribute>
      <xsl:attribute name="width">95 %</xsl:attribute>
      <xsl:attribute name="align">center</xsl:attribute>
      <xsl:element name="tr">
        <xsl:element name="td">
          <xsl:attribute
            name="align">center</xsl:attribute>
          <xsl:element name="table">
            <xsl:attribute
              name="width">100 %</xsl:attribute>
            <xsl:attribute
              name="style">BACKGROUND-
              COLOR: #ffbf2</xsl:attribute>
            <xsl:element name="tr">
              <xsl:element name="td">
                <h3>
                  <font face="Georgia">
                    <xsl:value-of
                      select="text()" />
                  </font>
                </h3>
              </xsl:element>
            </xsl:element>
          </xsl:element>
        </xsl:element>
      </xsl:element>
    </xsl:element>
  </xsl:template>
  <xsl:template match="//back/div/bibl">
    <xsl:element name="table">
  </xsl:template>
</xsl:stylesheet>

```

3.3. XSL Para los capítulos del libro

```

<?xml version="1.0" encoding="ISO-8859-15" ?>
<xsl:stylesheet
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="1.0">
  <xsl:output method="html" encoding="ISO-8859-15" />
  <xsl:template match="/">
    <html>
      <head>
        <title>Capítulo</title>
      </head>
      <body>
        <xsl:apply-templates select="capítulo/div" />
        <xsl:apply-templates select="capítulo/div/p" />
      </body>
    </html>
  </xsl:template>
  <xsl:template match="capítulo/div">
    <xsl:element name="table">
      <xsl:attribute name="border">0</xsl:attribute>
      <xsl:attribute name="width">95 %</xsl:attribute>
      <xsl:attribute name="align">center</xsl:attribute>
      <xsl:element name="tr">
        <xsl:element name="td">
          <xsl:attribute
            name="align">center</xsl:attribute>
          <xsl:element name="table">
            <xsl:attribute
              name="width">100 %</xsl:attribute>
            <xsl:attribute
              name="style">BACKGROUND-
                COLOR: #ffbf2</xsl:attribute>
            <xsl:element name="tr">
              <xsl:element name="td">
                <h3>
                  <font face="Georgia">
                    <xsl:value-of select="head/text()" />
                  </font>
                </h3>
              </xsl:element>
            </xsl:element>
          </xsl:element>
        </xsl:element>
      </xsl:element>
    </xsl:element>
  </xsl:template>
  <xsl:template match="capítulo/div/p">
    <xsl:param name="imagen" select="../figure/@entity" />
    <xsl:element name="table">

```

```

<xsl:attribute name="border">0</xsl:attribute>
<xsl:attribute name="width">95 %</xsl:attribute>
<xsl:attribute name="align">center</xsl:attribute>
<xsl:element name="tr">
  <xsl:element name="td">

    <div style="TEXT-ALIGN:justify">
      <font face="Georgia" size="2">
        <xsl:value-of select="text()" />
        <xsl:element name="sup">
          <xsl:element name="a">
            <xsl:attribute
              name="href">notas.html#Noteftn1
            </xsl:attribute>
            <xsl:attribute
              name="target">notes</xsl:attribute>
            <xsl:attribute
              name="class">notelink</xsl:attribute>
            <xsl:attribute name="title">Ir a la nota de
              pie</xsl:attribute>
          <xsl:element name="sup">
            <xsl:value-of select="."/emph / note / @id" />
          </xsl:element>
        </xsl:element>
      </xsl:element>
    <xsl:if test="figure / p / text() != ' '">
      <xsl:element name="image">
        <xsl:attribute
          name="src">
          <xsl:text>imagenes / </xsl:text>
          <xsl:value-of select="figure / @entity" />
        </xsl:attribute>
        <xsl:attribute
          name="align">baseline</xsl:attribute>
      </xsl:element>
      <p><xsl:value-of select="figure / p / text()" /> </p>
    </xsl:if>
  </font>
</div>
</xsl:element>
</xsl:element>
<xsl:element name="tr">
  <xsl:element name="td">
    <h2>
      <font name="Georgia" size="3">
        <xsl:value-of select="emph / text()" />
      </font>
    </h2>
  </xsl:element>
</xsl:element>
</xsl:element>
</xsl:template>
</xsl:stylesheet>

```

3.4. XSL para dar formato a la portada del libro

```

<?xml version="1.0" encoding="ISO-8859-15" ?>
<xsl:stylesheet
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="1.0">
  <xsl:output method="html" encoding="ISO-8859-15" />
  <xsl:template match="/">
    <html>
      <head>
        <title>menu</title>
      </head>
      <body>
        <table>
          <xsl:attribute
            name="width">50 % </xsl:attribute>
          <xsl:attribute
            name="align">center</xsl:attribute>
          <xsl:attribute
            name="height">100 % </xsl:attribute>
          <xsl:attribute name="style">BORDER-RIGHT: olive thin
            groove; BORDER-TOP: olive thin groove; Z-INDEX:
            101; LEFT: 8px; BORDER-LEFT: olive thin groove;
            BORDER-BOTTOM: olive thin groove;</xsl:attribute>
          <xsl:element name="tr">
            <xsl:element name="td">
              <xsl:attribute
                name="align">center</xsl:attribute>
              <xsl:apply-templates
                select="//front" />
            </xsl:element>
          </xsl:element>
        </table>
      </body>
    </html>
  </xsl:template>
  <xsl:template match="//front">
    <xsl:element name="table">
      <xsl:attribute name="bgColor">#ffbf2</xsl:attribute>
      <xsl:attribute name="border">0</xsl:attribute>
      <xsl:attribute name="width">100 % </xsl:attribute>
      <xsl:attribute name="align">right</xsl:attribute>
    <xsl:element name="tr">
      <xsl:element name="td">
        <xsl:attribute
          name="align">center</xsl:attribute>
      <xsl:element name="table">
        <xsl:attribute
          name="border">0</xsl:attribute>
        <xsl:attribute
          name="cellSpacing">0</xsl:attribute>
        <xsl:attribute
          name="cellPadding">0</xsl:attribute>

```

```

=> <xsl:element name="tr">
=>   <xsl:element name="td">
=>     <p align="center">
=>       <font name="verdana" size="5" color="#204060">
=>         <xsl:attribute
=>           name="width">2
=>           40</xsl:attribute>
=>       <STRONG>
=>         <xsl:value-of
=>           select="titlePage / docTitle / titlePart / emph / t
=>             ext()" />
=>         </STRONG>
=>       </font>
=>     </p>
=>   <p />
=> </p>
=> </xsl:element>
=> </xsl:element>
=> </xsl:element>
=> <xsl:element name="tr">
=>   <xsl:element name="td">
=>     <p />
=>   <p />
=> </xsl:element>
=> </xsl:element>
=> <xsl:element name="table">
=>   <xsl:attribute
=>     name="border">0</xsl:attribute>
=>   <xsl:element name="tr">
=>     <xsl:element name="tr">
=>       <xsl:attribute
=>         name="align">center</xsl:
=>         attribute>
=>       <xsl:element name="td">
=>         <font name="verdana" size="3">
=>           <xsl:attribute
=>             name="width">120</xsl:attribute>
=>         <b>
=>           <xsl:value-of
=>             select="titlePage / byline / docAuthor / text()"
=>           />
=>         </b>
=>       </font>
=>     </xsl:element>
=>   </xsl:element>
=>   <xsl:element name="tr">
=>     <xsl:attribute
=>       name="align">center</xsl:attribute>
=>     <xsl:element name="td">
=>       <font name="verdana"
=>         size="2">

```

```

        <xsl:attribute
            name="width">1
            20</xsl:attribute>
    = <b>
        <xsl:value-of
            select="titlePage / docEdition / emph / text()" />
        </b>
    </font>
</xsl:element>
</xsl:element>
= <xsl:element name="tr">
    <xsl:attribute
        name="align">center</xsl:attribute>
    = <xsl:element name="td">
        = <font name="verdana"
            size="2">
            <xsl:attribute
                name="width">1
                20</xsl:attribute>
            <b>
                <xsl:value-of
                    select="titlePage / docImprint / emph / text()" />
            </b>
            </font>
        </xsl:element>
    </xsl:element>
= <xsl:element name="tr">
    <xsl:attribute
        name="align">center</xsl:
        :attribute>
    = <xsl:element name="td">
        = <font name="verdana"
            size="2">
            <xsl:attribute name="width">120</xsl:attribute>
        = <b>
            <xsl:value-of
                select="titlePage / docDate / emph / text()" />
            <xsl:value-of
                select="titlePage / docDate / emph / lb / text
                ()" />
            </b>
            </font>
        </xsl:element>
    </xsl:element>
</xsl:element>
</xsl:element>
</xsl:element>
</xsl:element>
</xsl:element>
</xsl:template>
</xsl:stylesheet>

```

3.5. XSL para las notas de pie de página

```

<?xml version="1.0" encoding="ISO-8859-15" ?>
- <xsl:stylesheet
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  version="1.0">
  <xsl:output method="html" encoding="ISO-8859-15" />
  - <xsl:template match="/">
    - <html>
      - <head>
        <title>Notas</title>
      </head>
      - <body>
        <xsl:apply-templates select="notes/note" />
      </body>
    </html>
  </xsl:template>
  - <xsl:template match="notes/note">
    - <xsl:element name="table">
      <xsl:attribute name="border">0</xsl:attribute>
      <xsl:attribute name="width">95%</xsl:attribute>
      <xsl:attribute name="align">center</xsl:attribute>
      - <xsl:element name="tr">
        - <xsl:element name="td">
          <xsl:attribute
            name="align">center</xsl:attribute>
          - <xsl:element name="table">
            <xsl:attribute
              name="width">100%</xsl:attribute>
            >
            - <xsl:element name="tr">
              - <xsl:element name="td">
                - <font name="Tahoma"
                  size="2" color="blue">
                  <xsl:value-of
                    select="./@n" />
                </font>
              - <font name="Tahoma"
                size="2">
                  <xsl:value-of
                    select="./p/text()" />
                </font>
            </xsl:element>
          </xsl:element>
        </xsl:element>
      </xsl:element>
    </xsl:element>
  </xsl:template>
</xsl:stylesheet>

```

Anexo 4: Guía de usuario del Sistema Generador, módulo de ingreso de datos

En este anexo se describe detalladamente los procesos y la funcionalidad de los módulos del prototipo creado. Este es la primera parte del proyecto, módulo por el cual se ingresa los metadatos como: autor, materia, descripción y el contenido del libro para luego se procesarlo y obtener como resultado final el documento con el formato TEI que será indexado al base de archivos para luego ser recuperado por el sistema “Musuj Quipu UNMSM” que se detalla en el [Anexo 5].

4.1.- Creación de la carpeta

Para la creación de la carpeta de trabajo para el libro, de preferencia se debe asignar un código significativo ya que este será el identificador del libro en el sistema Musuj Quipu, por ello se recomienda colocar en nombre del autor y el inicial de sus apellidos, ejemplo:

Autor: BEATRIZ HERRERA GARCÍA

Identificado recomendado: beatriz_hg

Si ya existe un libro con un mismo identificador el sistema no permite crear, se tiene que variar en alguna letra el identificador, en caso de volúmenes siguientes de la misma obra se recomienda colocar el número del volumen al final, en la [Figura 4.1] se observa el interfaz por el cual se ingresa dicho código.

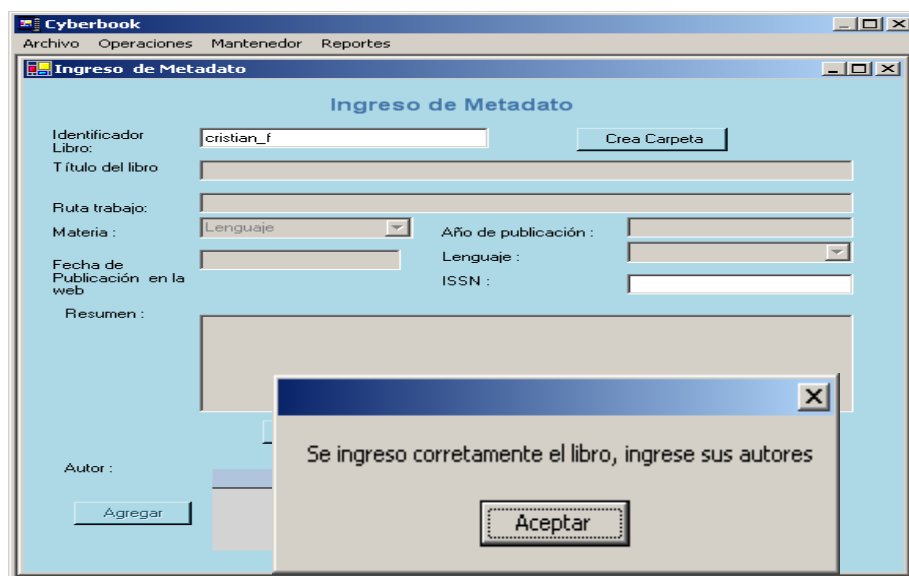


Figura 4.1: Asignación de código al libro y creando la carpeta de trabajo.

4.2.- Ingreso de datos

Si se creó correctamente la carpeta de trabajo, se habilitan los campos correspondientes al metadato del libro y deben ser ingresados, tal como se observa en la [Figura 4.2], siendo los datos a ingresar los que se detallan a continuación:

Título del libro: Ingresar el título completo del libro.

Ruta de trabajo: Es la ruta que anticipadamente se ha configurado, en donde se guardan todos los archivos correspondiente al libro que se está procesando.

Materia: Seleccionar la materia al cual pertenece el libro, si no existiera, se debe ingresar por otro interfaz.

Año de publicación: Año en cual fue publicado el libro físicamente.

Fecha publicación: Fecha en el cual se está publicando en el internet.

Lenguaje: Lenguaje en el cual está escrito el libro.

ISSN: Es el código perteneciente a cada libro físicamente.

Resumen: Ingresar breve resumen del contenido del libro.

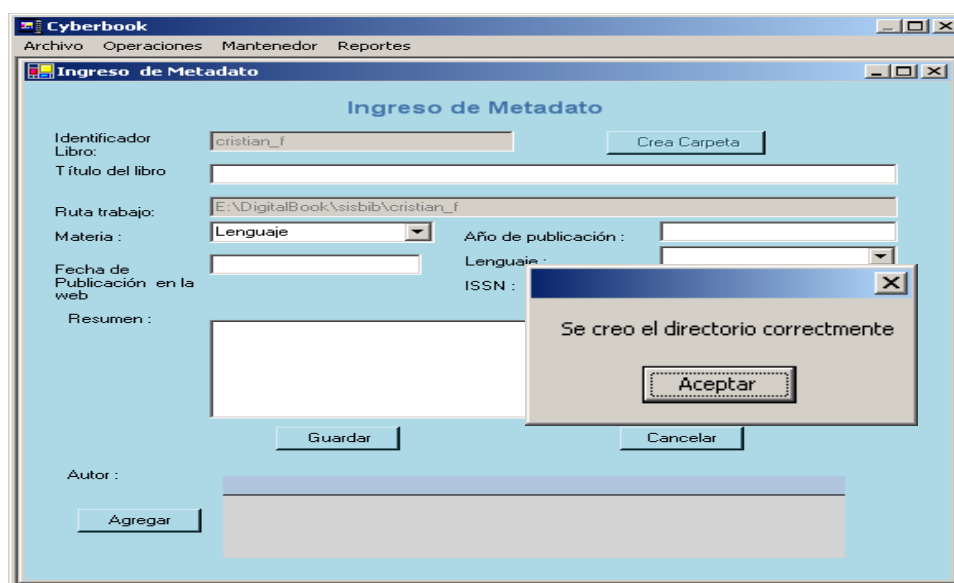


Figura 4.2: Interfaz para ingresar el metadato del libro.

Si el metadato se guardó correctamente se muestra el mensaje de información que se observa en la [Figura 4.3].

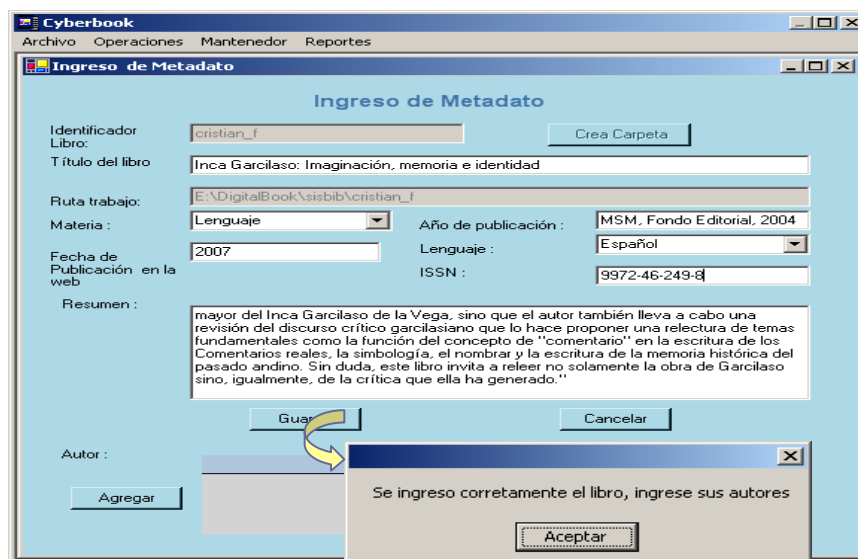


Figura 4.3 : Forma de ingresar el metadato del Libro.

4.3.- Ingreso y asignación de autor

Para agregar autor o autores al libro con el cual se está trabajando, hacer clic en el botón “Agregar” y se muestra el interfaz de ingreso de autor, en donde se puede realizar búsqueda por el apellido paterno, en caso de que el autor ya esté ingresado en el sistema, si no existe se debe ingresar sus datos requeridos en el interfaz que se observa en la [Figura 4.4].

idautor	nombre	apePaterno	apeMat
1	yanet	martinez	berro
2	yanet cilla	martinez	berro
3	yanet cilla	martinez	berro

Figura 4.4: Interfaz de ingreso de datos del autor

Una vez ubicado el autor que corresponde al libro, como se observa en la [Figura 4.5], agregar el autor a la obra haciendo clic en el botón “Agregar y Regresar” y a continuación se muestra el mensaje para regresar o seguir agregando más autores.

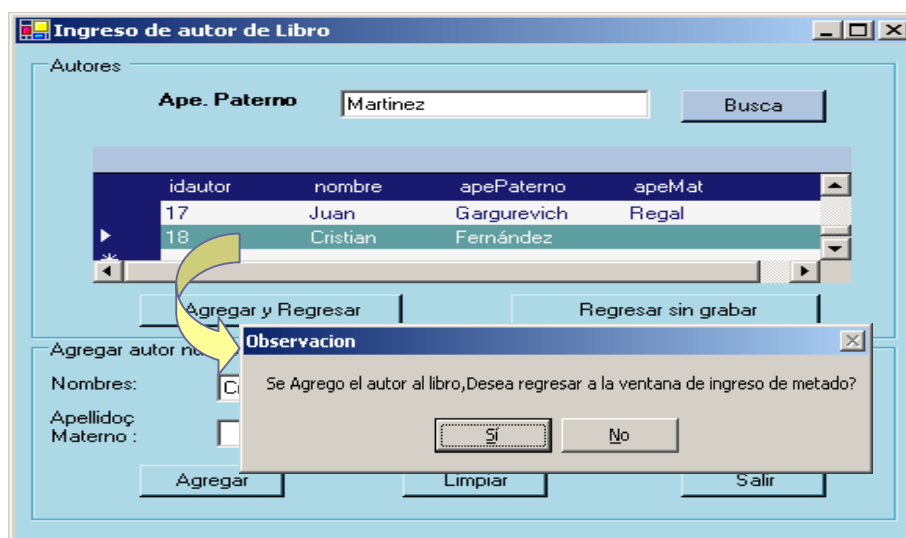


Figura 4.5: Asignando autor o autores al metadato del libro

Finalmente obtenemos el metadato del libro ingresado con todos los datos, tal como se observa en la [Figura 4.6].

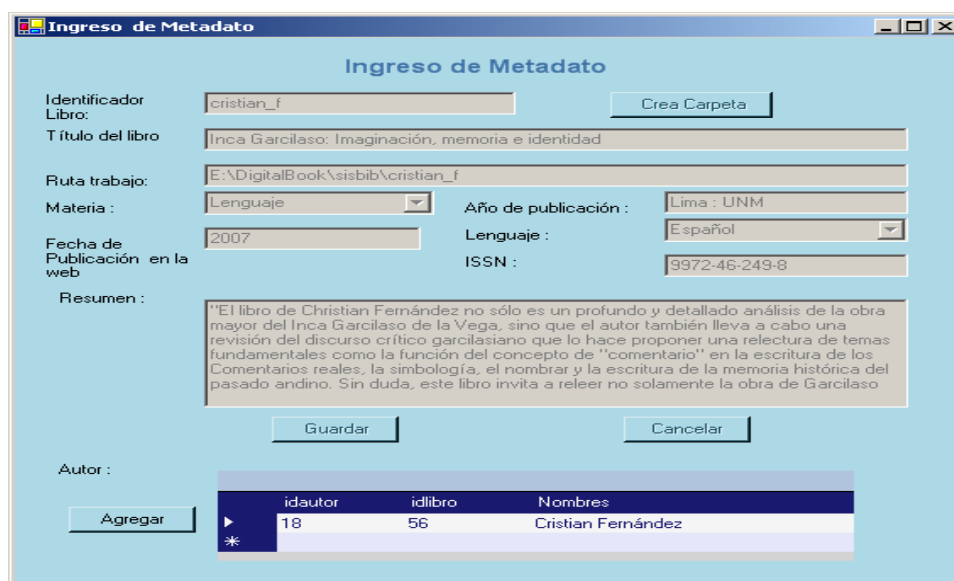


Figura 4.6: Metadato ingresado correspondiente a un libro.

4.4.- Modificación del metadato

Si se desea hacer alguna modificación del libro que fue ingresado con anterioridad, seleccione el módulo de “Lista de libros” en el cual se puede realizar la búsqueda del libro por el título del libro, no es necesario ingresar el título completo si no, solo basta ingresar una palabra o parte de ella contenida en el título del libro buscado y se obtendrá el libro para poder realizar los cambios necesarios o eliminarlo, el interfaz se muestra en la [Figura 4.7].

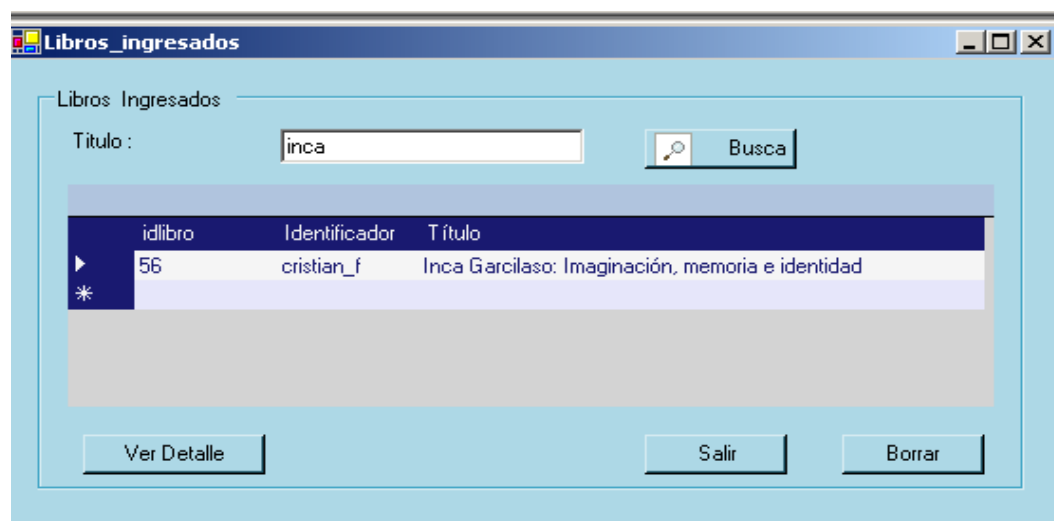


Figura 4.7: Listado de metadatos de libros ingresados.

4.5.- Preparación de carpetas y archivos

Luego de ingresar el metadato del libro, en la carpeta creada por el sistema de metadatos, crear una sub carpeta con el nombre “Recurso” y colocar dentro de ella el documento Word marcado, asignándole el mismo nombre que la carpeta creada, tal como se observa en la [Figura 4.8], además colocar dentro de ella el DTD del office.

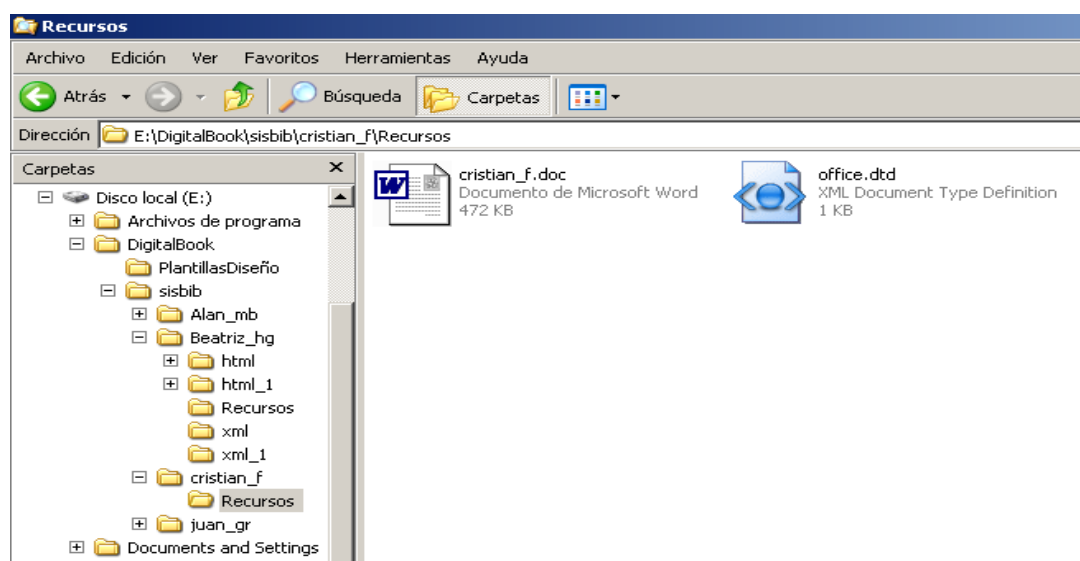


Figura 4.8: Contenido de la carpeta inicial de trabajo

4.6.- Generación del documento en el formato TEI y publicación

Seleccionar en el menú " Operaciones" el submenú "Generador" para ello ya se debe tener el documento en formato xx.SXW, perteneciente a al archivo de texto del openoffice, (abrir con el Openoffice el documento Word marcado y guardarlo en el formato del mismo SXW), este archivo se debe seleccionar tal como se observa en la [Figura 4.9].

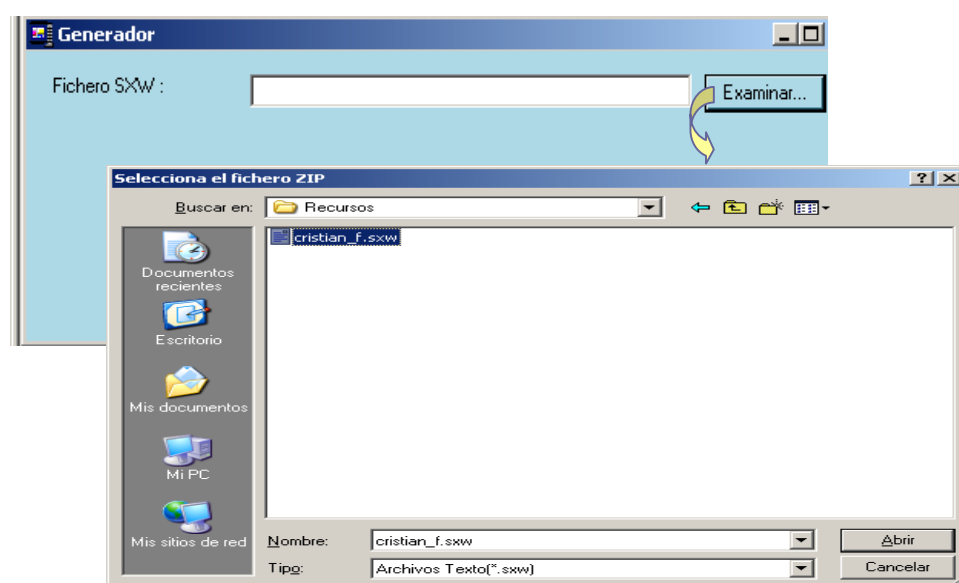


Figura 4.9: Seleccionando el archivo a transformar.

Una vez seleccionado el archivo, hacer clic en el botón “Empezar”, y el Sistema empezará a generar el documento tal como se observa en la [Figura 4.10], el sistema procesará el contenido del libro.

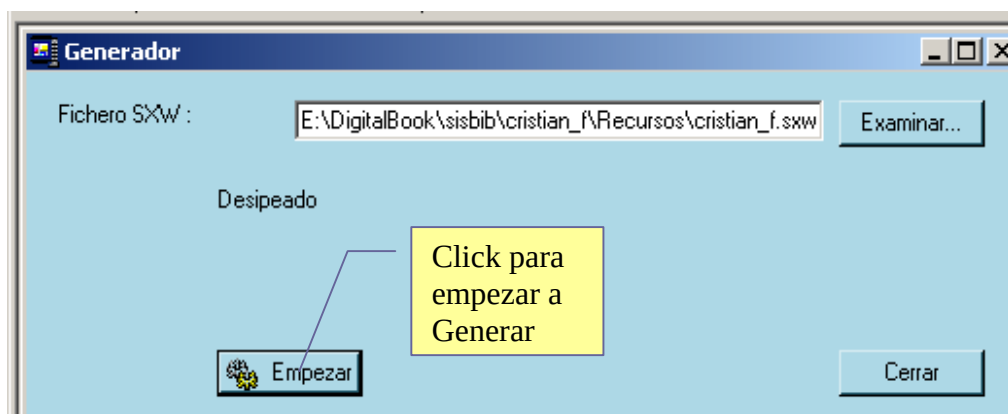


Figura 4.10: Procesando contenido del libro.

Finalmente se obtiene en libro ordenado, estructurado en formato XML y HMTL, ordenado cada uno en su respectiva carpeta. También el documento resultante se indexa para ser consultado mediante el buscador.

El sistema crea de manera automática la estructura que se observa en la [Figura 4.11] con todos los archivos correspondientes, la cantidad de los archivos que se muestra depende de el contenido de cada libro.

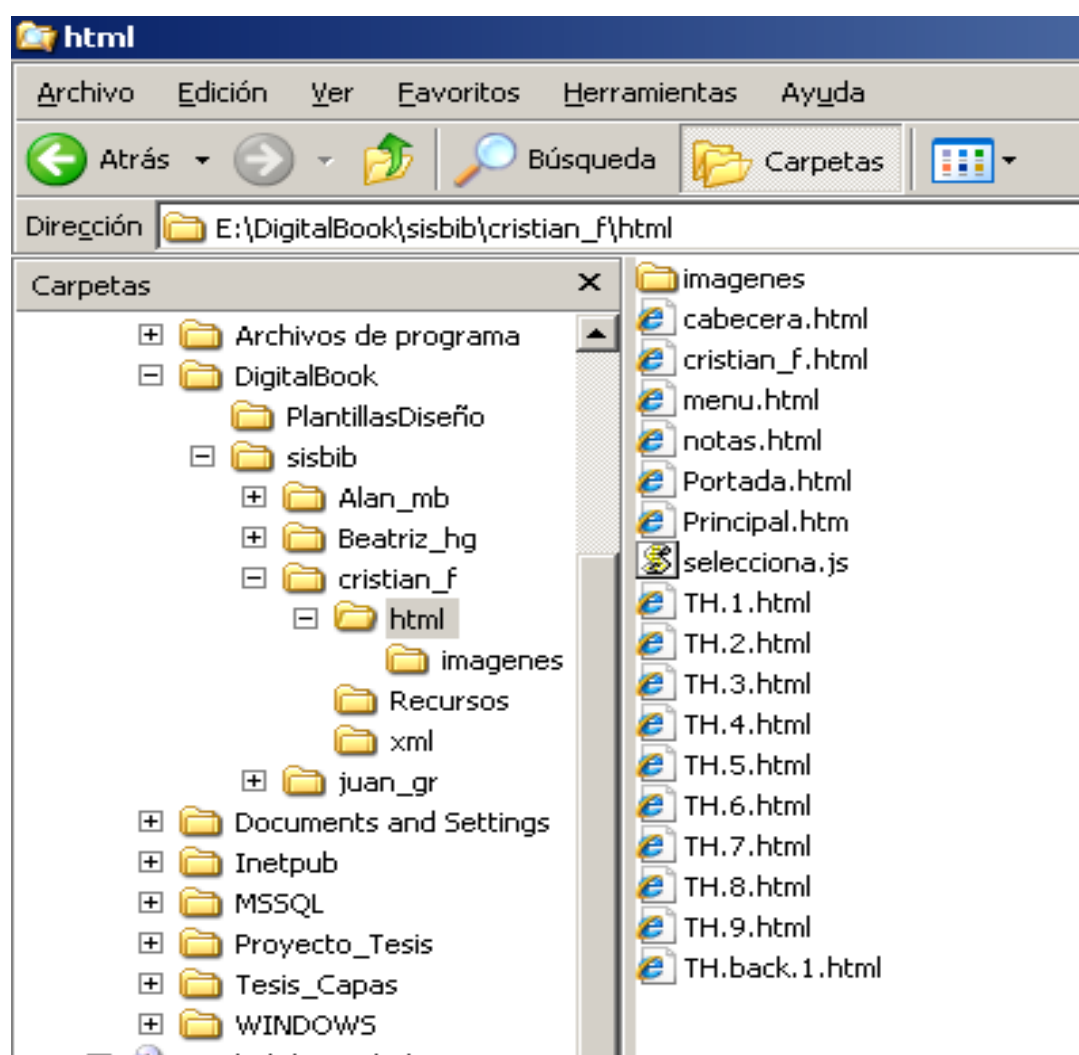


Figura 4.11: Estructura de un libro procesado mediante el modulo “Generador”

El archivo HTML con el nombre “Principal.htm” es la página principal que contiene el enlace a todas las partes del libro. Dicha página se muestra en la [Figura 4.12], donde se observa el libro completo y sus respectivos enlaces.

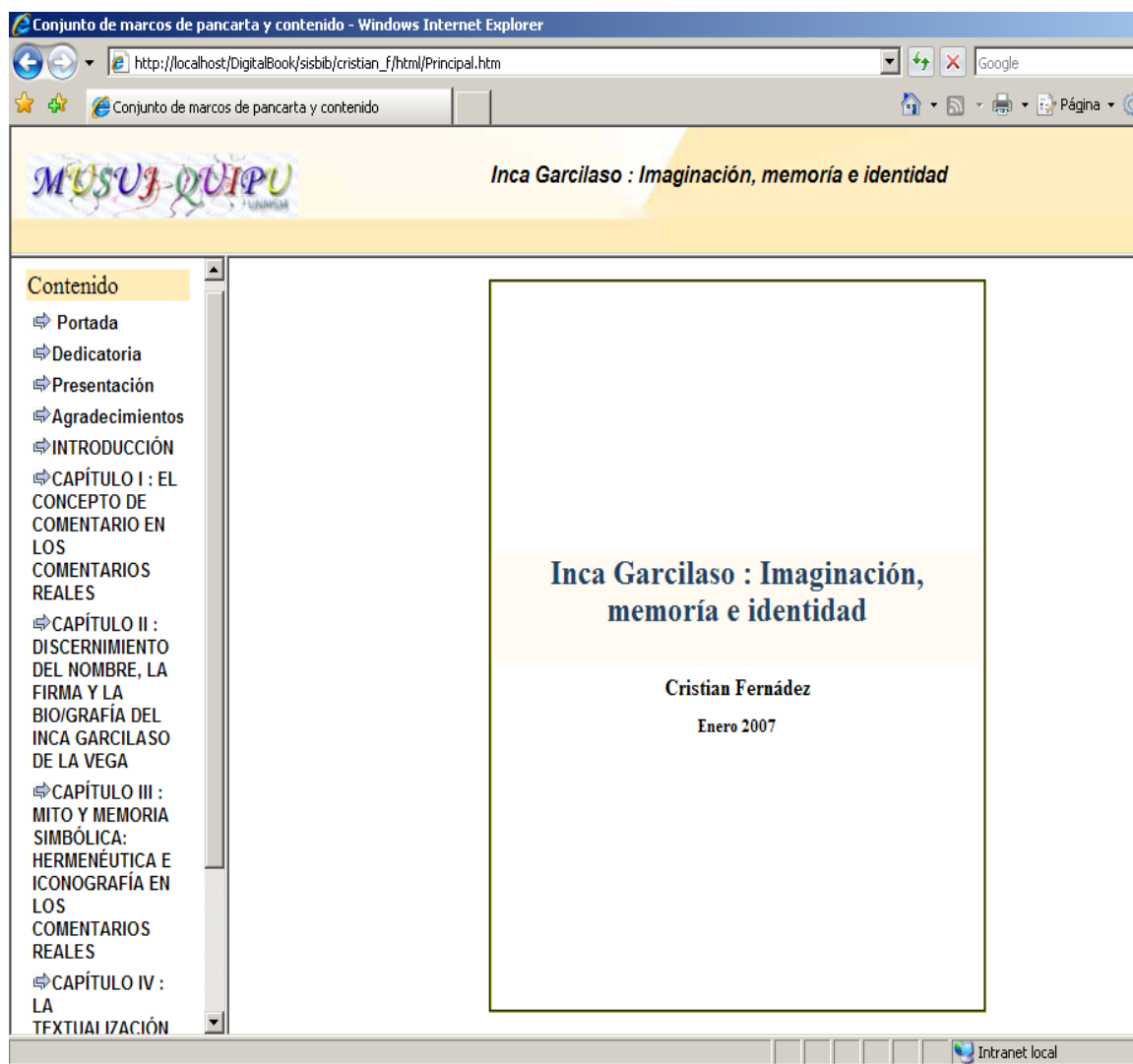


Figura 4.12: Resultado final del sistema, el contenido del libro en el formato HTML

Anexo 5: Guía de usuario del Sistema Musuj Quipu UNMSM

Musuj Quipu UNMS, es el Interfaz mediante el cual el usuario puede realizar sus búsquedas en el acervo bibliográfico que se tiene, vea la [Figura 5.1].

El usuario final solo puede visualizar este interfaz.

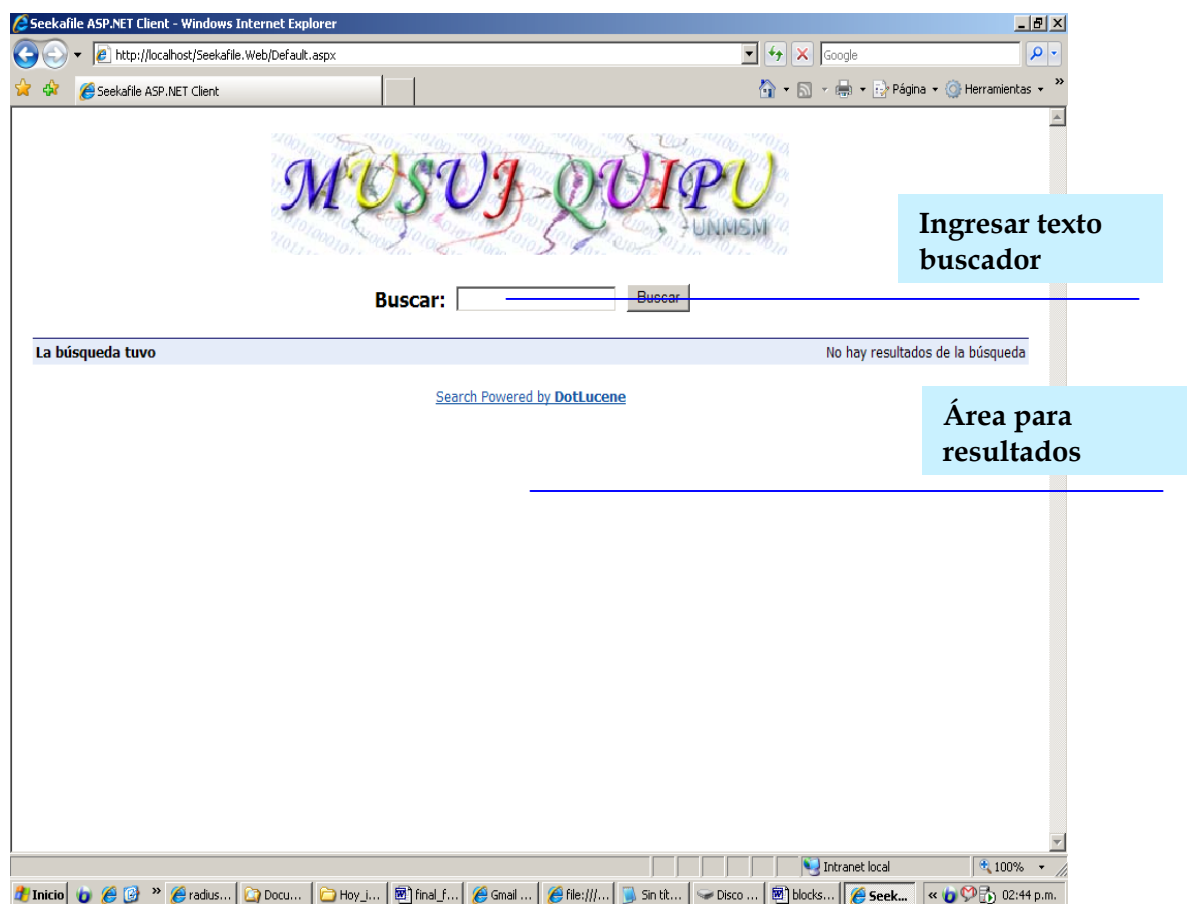


Figura 5.1: Interfaz para realizar búsquedas

Para realizar una búsqueda se debe ingresar el texto buscado y hacer click en el botón “Buscar”, se puede usar conectores como el “Y” y el “O” como se observa en la [figura 5.2]



Figura 5.2: Realizando Búsquedas

Los resultados de la búsqueda se listan como se observa en la [Figura5.3], en el cual se muestra detalladamente el listado de resultado.



Figura 5.3: Listado de resultados de la búsqueda realizada

En el listado de resultados se muestra lo siguiente:

- El autor o autores del libro.
- El título correspondiente al libro.
- Resumen del contenido en el cual se encuentra la palabra buscada y el metadato del libro.

Tal como se observa en la [Figura5.3]. Se tiene dos enlaces los cuales direccionan a las siguientes páginas que se observan la [Figura 5.4] y [Figura 5.5], donde se muestra el contenido del libro y su metadato respectivamente.

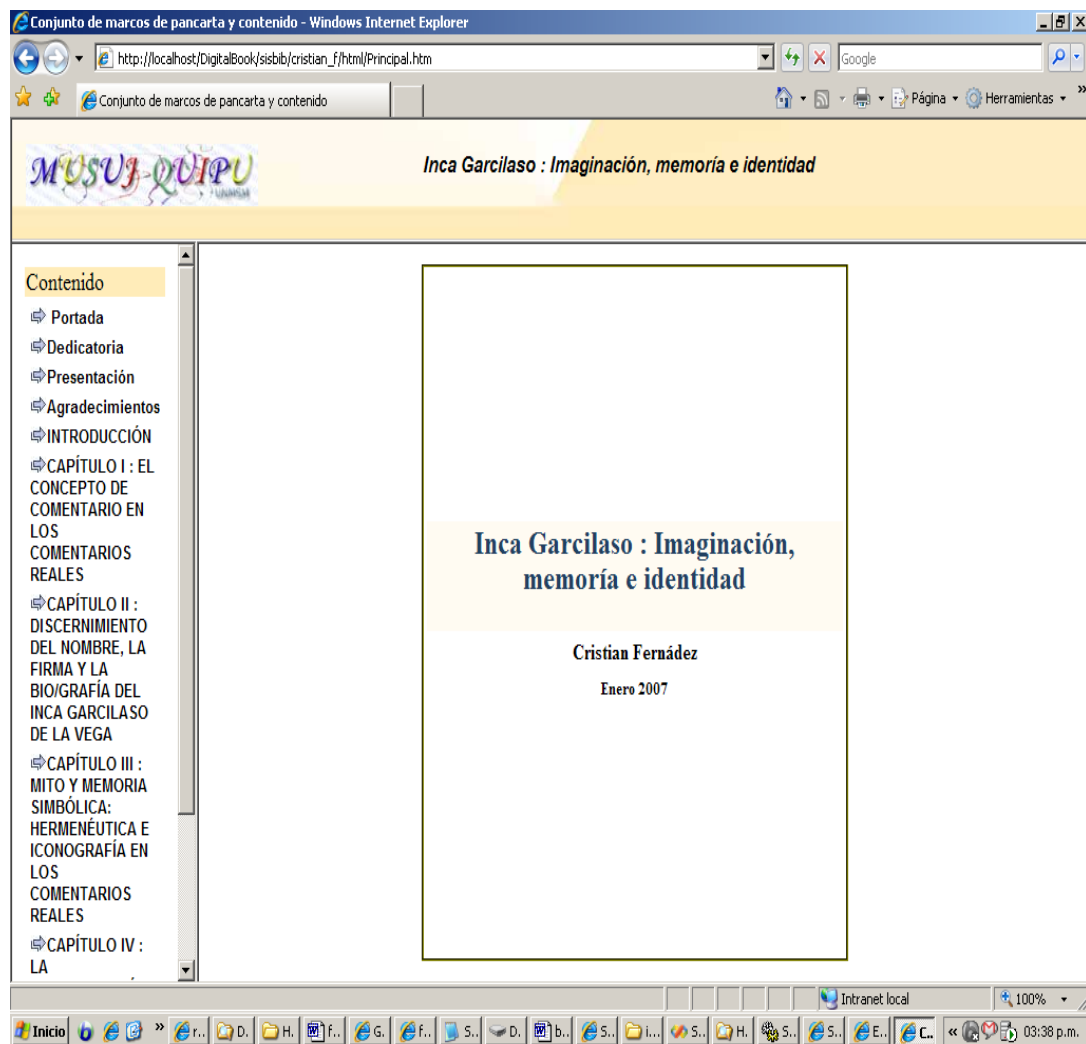


Figura 5.4: Contenido del libro seleccionado

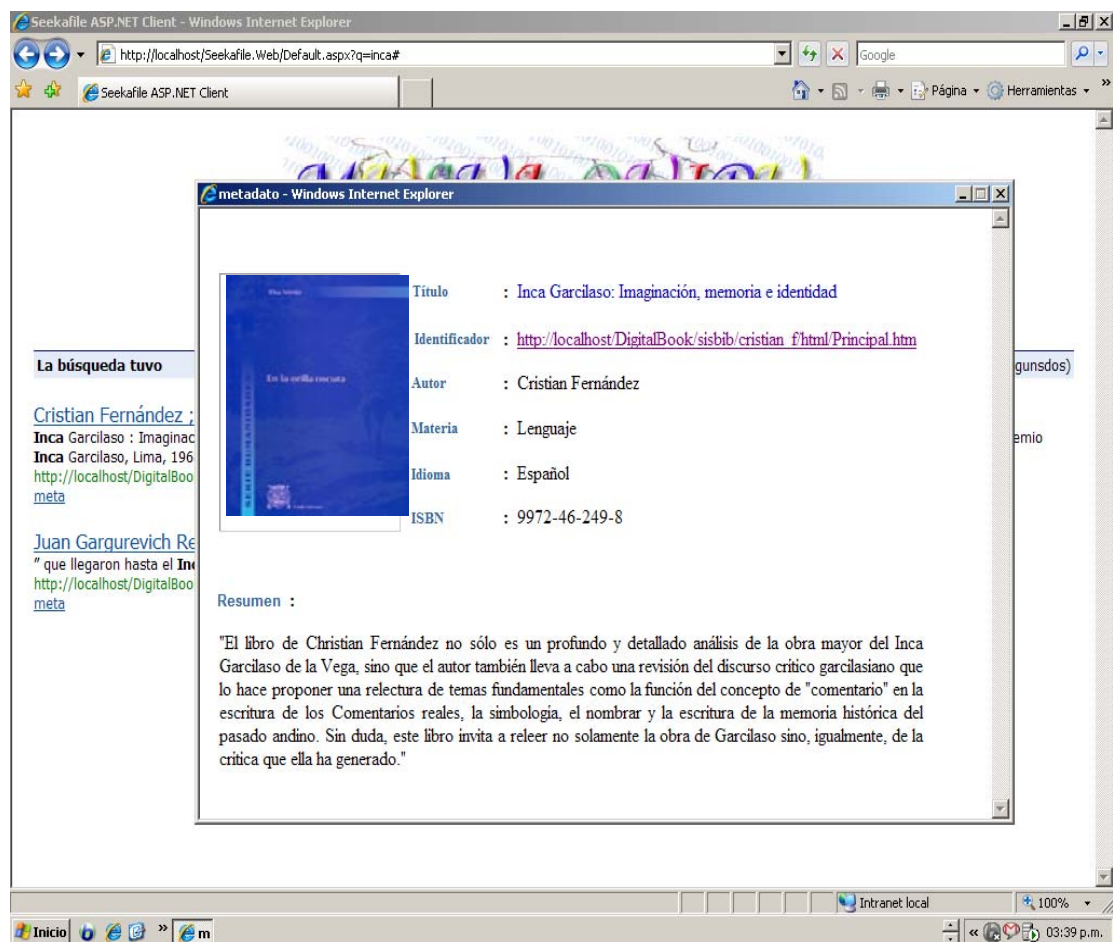


Figura 5.5: Contenido del libro seleccionado

En el metadato del libro seleccionado, se observa los datos ingresado con anterioridad en el Sistema Generador.