

Capítulo 3. Bibliotecas Digitales

En el capítulo I se mencionó la importancia y el auge que ha cobrado la investigación en bibliotecas digitales en los últimos años. En este capítulo, además de presentar algunas de las definiciones que existen sobre lo que es una biblioteca digital, se plantea el reto que conlleva la construcción de una biblioteca digital, así como las implicaciones físicas, tecnológicas e incluso económicas que implica esta nueva tecnología.

También se describen en este capítulo algunas arquitecturas propuestas por diversos investigadores del área. Se presentan algunos ejemplos de universidades y centros de investigación que se han enfrentado al desafío de crear una biblioteca digital. Se hace énfasis en aquellos trabajos que incorporan la tecnología de agentes en la creación de bibliotecas digitales, puesto que es en este contexto que se desarrolla la presente tesis.

Finalmente, en la última sección se describen los principales algoritmos de recuperación de información que son usados en el área de bibliotecas digitales.

3.1 ¿QUÉ ES UNA BIBLIOTECA DIGITAL?

La frase "biblioteca digital" provoca una impresión diferente en cada lector. Para algunos es simplemente la automatización de bibliotecas tradicionales. Para otros es una colección de manuscritos digitalizados. Para muchos profesionales de la computación una biblioteca digital es simplemente un sistema de información distribuido basado en texto; una colección de servicios de información distribuidos, un espacio distribuido de interligas de información, o un sistema de información multimedios [Fox et al. 1993]. Aquellos que estudian tecnologías de colaboración ven a las bibliotecas digitales como el espacio en el cual la gente se comunica, comparte y produce nuevo conocimiento. Aquellos que trabajan en tecnología de la educación ven a las bibliotecas digitales como el soporte para el aprendizaje, tanto formal como informal [Iarchionini y Maurer 1995].

Birmingham usa el término de biblioteca digital como un nombre genérico para estructuras dinámicas federadas que proporcionan acceso intelectual y físico a redes crecientes de información codificada en formatos digitales

multimedios [Birmingham 1995].

Desde una perspectiva de bases de datos o recuperación de información, las bibliotecas digitales pueden ser vistas como una forma de bases de datos federadas [Nurnberg et al. 1995].

Desde una perspectiva de hipertextos el campo de bibliotecas digitales podría ser una aplicación particular de la tecnología de hipertexto.

Para [Graham 1995] una biblioteca digital es una colección de información electrónica organizada. Las bibliotecas digitales son colecciones de información existentes en varios lugares y accesibles a través de varias herramientas disponibles.

Fox y Lunnin definen una biblioteca digital como una tecnología de ambiente distribuido la cual reduce dramáticamente las barreras de la creación, disseminación, manipulación, almacenamiento, integración y reuso de la información por individuos y grupos [Fox y Lunnin 1993].

Las definiciones son diversas y varían según el punto de vista del usuario pero la característica en común entre todas ellas es que la biblioteca digital debe contener los servicios esenciales de las bibliotecas tradicionales y además debe explotar las ventajas de almacenamiento, búsqueda y comunicación digital. Para lograr este objetivo, se deben superar los retos que se plantean en la siguiente sección.

3.2 RETOS EN LA CONSTRUCCIÓN DE BIBLIOTECAS DIGITALES

Las bibliotecas digitales no residen en una construcción física, ni manejan objetos físicos como libros, catálogos o archiveros. De hecho las bibliotecas digitales, deben trasladar los elementos físicos y conceptuales de la biblioteca física y adaptarlos al ambiente digital e incluso se deben añadir nuevos elementos digitales que no aparecen en la biblioteca física [Garret 1993]. Esto nos conduce a una serie de problemas que deben ser resueltos para lograr crear una verdadera biblioteca digital.

Para la construcción de una biblioteca digital se deben tomar en cuenta los siguientes aspectos [Birmingham 1995].

La variedad de necesidades del usuario.

Los tipos heterogéneos de recursos de información.

Los tipos de datos multimedios.

Las nuevas colecciones de datos.

La posible sobrecarga en la información.

El tipo de herramientas, que se le deben proporcionar al usuario para que la recuperación de información se simplifique. El desarrollo de un mecanismo de acceso efectivo que asegure a los dueños y usuarios que la información esta protegida contra accesos no autorizados, ya sea accidentales o intencionales.

Las técnicas de colección, almacenamiento y organización de la información que deben existir en toda biblioteca digital, así como los métodos de búsqueda, selección y recuperación de información que se deben implementar para funcionar sobre redes de cómputo distribuido.

Además de tener en mente las necesidades ya mencionadas, es necesario cuestionarse en cuanto a: ¿cómo se comunicarán entre bibliotecas? ¿cómo sabrá el usuario donde buscar? ¿cómo deben ser ordenados, seleccionados, unidos y transmitidos los resultados al usuario? ¿cómo asegurar la integridad y seguridad de los datos? ¿cómo digitalizar los objetos y ponerlos en línea? ¿cómo encontrar fallas y corregirlas? ¿cómo actualizar la información? ¿cuánto costará la infraestructura de comunicaciones? ¿cuánto costará la infraestructura de cómputo? ¿cuánto costará la información?

Kuny [1996] plantea el problema de la preservación, según este autor para preservar la información es necesario que las bibliotecas digitales migren continuamente sus datos. Pero el cambio continuo en los métodos para grabar información, en los formatos de almacenamiento y en las tecnologías que se usan para mantener la información, hace que el costo de la migración sea elevadísimo e incluso incalculable. Sin embargo la función de preservación debe ser atendida en todas las colecciones digitales. Es necesario "refrescarse" y migrar hacia nuevos sistemas para mantener la accesibilidad a la información

Otro problema que se plantea es el derecho de autor. Según Kuny, las cuestiones de derecho de autor deben ser contempladas para que la información pueda ser creada y distribuida a través de bibliotecas digitales de una manera equitativa tanto para el productor como para el consumidor de dicha información.

Otro aspecto importante para este autor es el costo, la información nunca ha sido realmente gratis ya que siempre existe un costo en su creación, su producción y su diseminación. En el caso de bibliotecas públicas subsidiadas por el gobierno, el costo real de la información esta oculto, pero la creación de bibliotecas digitales introduce una nueva e incierta realidad económica. La era digital requiere que los clientes paguen por acceder a los servicios y colecciones digitales.

Con respecto a la obtención de información en bibliotecas digitales. Kuny afirma que se requiere de una amplia investigación en la adquisición de

conocimiento, para lograr interpretar adecuadamente los *queries* (*preguntas*) del usuario. Además menciona que se requiere investigación sobre la extracción dinámica de llaves de información pertinentes para los *queries* del usuario, para lograr devolver un resumen de contenido que facilite la retroalimentación del usuario, en vez de simplemente proporcionar un resumen libre de contexto.

Para Griffiths y Kertis [1990], una representación apropiada de los recursos de información conduce a una identificación y selección exitosa de los mismos. Esto es importante debido a que en el ambiente de red se cuenta con un gran número de recursos digitales que contienen muchas formas de información (texto, imágenes, sonido, software, estructuras de datos complejos, etc.), así que es difícil para los usuarios conocer que recursos están disponibles. Por lo tanto es necesario extraer la información específica que ellos necesitan desde fuentes apropiadas.

3.3 COMPONENTES DE UNA BIBLIOTECA DIGITAL

El problema central con la traducción de datos físicos es decidir que aspectos deben tomarse en consideración durante el proceso de traducción. Esto no es sencillo, ya que las características físicas pueden representar algo para unos personas y otra cosa muy diferente para otras. El problema con la traducción de metadatos de la biblioteca física es que a veces los metadatos o su aplicación están influenciados por el dato físico. El problema clave con la traducción de procesos de la biblioteca física es que son ejecutados por humanos y muchas veces los procesos de ejecución son informales.

Otras arquitecturas se encuentran detalladas en [Shen et al 1994] y en [Srihari et al. 1994].

Finalmente y a pesar de las diversas arquitectura propuestas, el objetivo de toda biblioteca digital es servir de mediadora entre los clientes, bibliotecarios y editores y el sistema de computo.

Tabla 3.1 Taxonomía de elementos en una biblioteca digital (Adaptada de [Nurnberg et al 1995]).

	Datos	Metadatos	Procesos
Traducción de entradas de	Libro	Indice estático	Adquisición de datos

una biblioteca física	Revista Video	Clasificaciones Ordenamiento espacial	Sugerir fuentes Ayudar a localizar fuentes
Nuevas entidades de una biblioteca digital	Hypernovela Visualización científica Programa	Indice dinámico Estructura personalizada Anotaciones	Búsqueda de textos Presentación personalizada Recuperación por agentes

3.4 ALGUNAS BIBLIOTECAS DIGITALES

Desde el artículo "As we may think" de Vannevar Bush en 1945 hasta la fecha, mucho se ha dicho, escrito y hecho en materia de bibliotecas digitales. En la actualidad se cuenta con diversos proyectos de bibliotecas digitales con muy variados objetivos, entre los que podemos mencionar:

En un proyecto IBM/Universidad Reserve Western, la biblioteca tiene la intención de ser una componente de un "ambiente de aprendizaje electrónico", como parte de un sistema de información [Backer y Kingman 1993].

En el proyecto de la Universidad Intermedia Brown, el objetivo es concebir repositorios para materiales interactivos de instrucción [Yankelovich et al. 1988].

El proyecto Sequoia pretende crear una herramienta específica para el manejo de datos que componen una investigación científica masiva [Kochevar et al. 1993, Stonebraker y Dozier 1991].

El proyecto de la Biblioteca de Memoria del Congreso Americano, intenta ampliar el acceso público a material histórico [Culshaw 1992, Polly y Lyon 1992].

Un proyecto de la Universidad Cornell auspiciado por la Comisión de Preservación, intenta rescatar el contenido de materiales frágiles [Kenney y Personius 1993; Webster 1993].

El proyecto Envision del Instituto Politécnico de Virginia se enfoca a la recuperación de información e interfaces de objetos estructurados y reusables, así como a tecnologías para aprendizaje acerca de la ciencia computacional con material de ACM, IEEE-CS y otros [Fox et al. 1993].

Un proyecto conjunto de la Biblioteca del Vaticano, la Universidad Pontificia Católica de Rio de Janeiro e IBM pretender dar acceso desde WWW a los materiales históricamente significativos así como a hermosas ilustraciones copiadas de la Divina Comedia de Dante y a los cuatro viejos manuscritos de los poemas de Virgilio [Gladney et al. 1994].

El proyecto Nacional de Bibliotecas Digitales (NDLP) de la Biblioteca del Congreso pretende construir un sistema experimental para organizar y almacenar colecciones históricas. Se van a convertir las colecciones históricas a forma digital para hacerlas disponibles en Internet. Esto incluye texto, fotos e ilustraciones [Arms et al. 1997].

La Biblioteca Digital de imágenes de Astronomía (AD IL) va a coleccionar imágenes astronómicas para hacerlas disponibles a la comunidad astronómica y al público en general. Los clientes accesan a través de Internet usando cualquier navegador multimedios para buscar imágenes. Una vez localizadas dichas imágenes, los usuarios pueden salvarlas en sus computadoras locales para posteriormente analizarlas [Ye 1997]

El proyecto CoLib es un esfuerzo multi-institucional, inter e intradisciplinario que pretende establecer a gran escala una biblioteca digital botánica distribuida. El proyecto engloba a cinco proyectos particulares que son: el proyecto de la Flora de Norte América (FNA), el proyecto de la Flora de Peru, el proyecto de la Flora de Madagascar, el proyecto de La Flora Mesoamericana y el proyecto de la Flora de China Los mayores participantes en el proyecto CoLib son la Universidad de Washington, Texas A&M University, y el Jardín Botánico de Missouri. A raíz del proyecto se han creado diversas líneas de investigación, en las siguientes áreas: (1) sistemas de bibliotecas colaborativas hipermedios, (2) cómputo multimedios interactivo y sistemas de comunicaciones, y (3) relaciones bibliotecarias en la biblioteca digital botánica [Schnase et al. 1994].

Estos ejemplos muestran claramente la importancia que ha cobrado el desarrollo de bibliotecas digitales.

En particular hay una iniciativa para bibliotecas digitales (DLI) que está conjuntamente financiada por la Fundación Nacional de Ciencias (NSF), la Agencia de Proyectos de Investigación Avanzada del Departamento de Defensa (DARPA) y la Administración Nacional de Aeronáutica y del Espacio (NASA). Este proyecto de investigación se originó a raíz de la creciente demanda por sistemas y servicios, capaces de proveer a las diversas comunidades de usuarios, con un acceso coherente a bases de datos extensas y geográficamente dispersas.

El objetivo de la iniciativa es incrementar dramáticamente los medios de obtener, almacenar y organizar información en forma digital y hacerla disponible para la búsqueda, recolección y procesamiento a través de redes de comunicación. Un aspecto importante de la iniciativa es el de establecer mejores conexiones entre ciencia básica e innovación tecnológica.

La investigación sobre bibliotecas digitales explora temas variados que son relevantes para el desarrollo y uso de colecciones bibliográficas distribuidas electrónicamente. Las colecciones de interés pueden contener texto, imágenes, mapas, registros de audio, video y películas, además de combinaciones de estos medios (multimedios). La investigación sobre bibliotecas digitales se concentra también en temas sociales, económicos y culturales asociados con el desarrollo y uso de estos recursos.

Seis proyectos multidisciplinarios están siendo financiados a lo largo de un periodo de cuatro años. Cada proyecto tiene equipos de investigadores de una universidad líder con investigadores de otras organizaciones. Más de 75 organizaciones diferentes han establecido relaciones de asociación con los proyectos. Las organizaciones asociadas representan diversos intereses e incluyen empresas líderes de Estados Unidos en los rubros de computación y comunicaciones, instituciones académicas en todos los niveles, bibliotecas, editoriales, gobierno y agencias estatales, asociaciones profesionales y otras organizaciones interesadas en bases de datos y manejo de información a gran escala.

Universidad de California Berkeley. Planeamiento ambiental y sistemas de información geográfica. Este proyecto producirá un prototipo de biblioteca digital enfocado a la información ambiental. La meta del proyecto es desarrollar tecnologías para el acceso inteligente a colecciones distribuidas de fotografías, imágenes de satélite, mapas, documentos completos y documentos multivalentes. El trabajo está concentrado en la recuperación de información geográfica (GIR) [Larson 1996]. Coleccionará información diversa sobre el ambiente para la preparación y evaluación de datos ambientales, reportes de impacto ambiental y material relacionado. La investigación pretende crear sistemas de producción CERES, para crear el prototipo los investigadores necesitarán producir tecnologías que permitan que usuarios no entrenados contribuyan para hallar información relevante en otros sistemas de biblioteca digital en el WWW. Las áreas de investigación incluyen indexado automático, recuperación inteligente, procesos de búsqueda, tecnología de bases de datos para soportar aplicación de bibliotecas digitales, nuevas técnicas para análisis de documentos, y compresión de datos y herramientas de comunicación para soportar aplicaciones de bibliotecas digitales. Participantes y colaboradores en el proyecto incluyen a Xerox Corporation, Agencia de recursos de California, Biblioteca Estatal de California, Biblioteca del condado de Sonoma,

Asociación de gobiernos de San Diego, Corporación The Plumas, Oficina de Educación del condado de Shasta, Hewlett Packard.

Universidad de California en Santa Bárbara. Proyecto Alejandria: Información geográfica espacialmente referenciada. El proyecto de Alejandria desarrollara una biblioteca digital que proveerá de un acceso fácil a grandes y diversas colecciones de mapas, imágenes y materiales pictóricos, así como a un amplio rango de nuevos servicios electrónicos de biblioteca [Frew et al. 1996]. El proyecto está centrado en la Universidad de California, Santa Bárbara, la cual tiene la mayor colección de mapas e imágenes y un fuerte grupo de investigación enfocado al área de información espacialmente indexada. Este proyecto involucra a la Universidad Estatal de Nueva York en Buffalo (SUNY-Buffalo), la Universidad de Maine y varios patrocinadores industriales. El proyecto comenzará con colecciones de mapas digitalizados, imágenes y fotos aéreas relacionados con los condados de Santa Bárbara, Ventura y Los Angeles usando software desarrollado para sistemas de información geográfica. A lo largo de 4 años, el proyecto crecerá para incluir otros componentes de UCSB, SUNY-Buffalo, Biblioteca del Congreso, Survey Geológico de los Estados Unidos y la Biblioteca Pública de San Louis, así como otras bibliotecas interesadas. Cada sitio incluirá facilidades para interfaces de información geográfica, catálogos electrónicos, y almacenamiento y adquisición de información.

Universidad Carnegie Mellon. Biblioteca de video digital Informedia. Habilitará a los usuarios para acceder, explorar y recuperar material de ciencia y matemáticas desde archivos de video [Christel 1995]. El sistema Informedia trabaja con integración de pláticas, imágenes y tecnologías para entendimiento de lenguaje natural. La biblioteca inicialmente contendrá 1000 horas de video. Una biblioteca digital de videos debe ser eficiente para dar a los usuarios precisamente el material que ellos necesitan, para hacer recuperación rápida, la biblioteca digital de video necesitará soportar particionar el video en pequeños clips y representaciones alternativas del video. Aspectos como interacción hombre-computadora, precio y cargo por uso de video digital, y seguridad y privacidad, son las líneas a seguir como parte del programa de investigación. Participantes industriales incluyen a Microsoft Corporation, Digital Equipment Corporation, Bell Atlantic Network Services, Comunicaciones QED, Universidad Abierta, y Escuelas Públicas del condado de Fairfax Va.

Universidad de Illinois de Urbana-Champaign. Compiladores de literatura científica federados. Este proyecto está a cargo del nuevo Centro de Información de Bibliotecas en Ingeniería Grainger en la Universidad de Illinois en Urbana-Champaign y estará centrado en revistas y artículos en ingeniería y literatura científica [Schatz et al. 1996]. Un sistema prototipo incluirá una interfaz de usuario basado en una versión personalizada de Mosaic, software desarrollado por la universidad patrocinado por NSF

para ayudar a los usuarios a navegar en el WWW. Una vez que la biblioteca digital este desarrollada, técnicas de recuperación serán creadas, usando patrones del sistema analizado, y los efectos de los costos de la información serán determinados. Participantes en el proyecto incluyen al NCSA, la Universidad de Arizona, IEEE, Sociedad Americana de Física, John Wiley & Sons, US News & World Report. Actualmente se esta desarrollando una infraestructura de información para realizar búsqueda efectiva de documentos técnicos en el Internet. Se están construyendo repositorios de colecciones indexadas de multiples formas para buscar el material vía multiples vistas desde una colección virtual. Primero se empieza con una entrada típica a la biblioteca digital que es un término específico a buscar, después se buscan los documentos que coincidan con el término dado. El usuario puede desplegar los documentos encontrados en diferentes niveles de detalle para generar búsquedas subsecuentes hasta que gradualmente recupere documentos realmente relevantes para él.

Universidad de Michigan. Agentes Inteligentes para la localización de la información. El proyecto de la Biblioteca Digital de la Universidad de Michigan está investigando métodos para lograr acceso a fuentes de información distribuidas y heterogéneas. Los accesos deben ser independientes del tiempo y la distancia, flexibles y personalizados para cada individuo a través de una arquitectura federada y distribuida usando agentes que tengan conocimiento sobre las colecciones, los usuarios y los métodos de procesamiento de términos a buscas, así como procedimientos de mediación para coordinar interacción entre ellos [Wellman et al. 1996]. UMDL será organizado como un sistema distribuido consistente de fuentes de información, interfaces de usuario y un conjunto de agentes de procesamiento [Atkins et al. 1996]. El proyecto conducirá investigación coordinada para crear, operar, usar y evaluar de una biblioteca digital multimedia que esta en continua evolución. El contenido de la biblioteca será ciencias del espacio y de la sierra. Potencialmente estarán conectados miles de usuarios y de repositorios de información, el sistema de biblioteca será diseñado para conocer las necesidades de sistematizar la gran cantidad de información desde un arreglo de tópicos disponibles en Internet. Un componente crítico del proyecto es la prueba y evaluación del sistema prototipo por una amplia variedad de usuarios, incluyendo aquellos de bibliotecas publicas, y de escuelas locales. Patrocinadores comerciales incluyen a IBM, Elsevier Science, Computadoras Apple, Bellcore, UMI internacional, Mc-Graw-Hill, Enciclopedia Britanica y Kodak.

Universidad de Stanford. Mecanismos interoperacionales entre servicios heterogéneos. El proyecto de la Biblioteca Digital Integrada de Stanford desarrollará tecnologías para integrar una biblioteca "virtual" que proporcione acceso uniforme a colecciones de información digital. La biblioteca digital integrada creará un ambiente compartido que ligara desde colecciones personales de información, las colecciones halladas

hoy en bibliotecas convencionales, hasta la gran cantidad de colecciones de datos compartidos por científicos. Se pretende crear tecnología que permita a los usuarios acceder a los recursos digitales desde documentos estáticos para la exploración dinámica de la información y el manejo de servicios, sin tener que conocer los detalles del formato de cada documento y los mecanismos de cada servicio. Esta tecnología es llamada InfoBus, la cual proporciona un marco de trabajo unificado que reúne los servicios proporcionados por el WWW, así como los servicios de recuperación de información y nuevos tipos de servicios que están siendo desarrollados. La investigación incluirá: modelos de comunicación, interfaces de información para clientes, y servicios para hallar información. Las organizaciones participantes incluyen a ACM, Bellcore, Dialog, EIT, Laboratorios Hewlett Packard, ITC, O Reilly y asociados, NASA Ames, WAIS Inc., y PARC. El protocolo "InfoBus", proporciona una manera uniforme para acceder una variedad de servicios y fuentes de información a través de "proxies" que actúan como intérprete entre el protocolo InfoBus y el protocolo nativo [Paepcke 1996]. InfoBus está acompañado por una interfaz que proporciona un modelo consistente para los usuarios. Se trabaja además en una interfaz de usuario llamada DLITE [Cousins 1997], la cual permite personalizar las tareas del usuario, DLITE esta diseñada para facilitar a los usuarios interactuar con los diferentes servicios mientras se enfoca en una tarea específica. DLITE introduce el concepto de centros de trabajo (workcenter) que es un lugar donde las herramientas para una tarea están listas para usarse. Las componentes de DLITE caen en 5 categorías básicas: documentos, colecciones, *queries*, servicios y representaciones de la gente. Usa la metáfora de workcenters conteniendo componentes y mecanismos de retroalimentación como animación, en el contexto del prototipo de una biblioteca digital

Además de las interfaces de usuario, una de las áreas de investigación con mayor empuje en el proyecto de la Biblioteca Digital de Stanford es ayudar a los usuarios a hallar información. Se han iniciado una serie de proyectos en este tema. Se han buscado maneras para que las herramientas de búsqueda puedan ser usadas a través de múltiples fuentes que usan diferentes sintaxis o lenguajes. También se han desarrollado herramientas que proporcionan filtrado estadístico o colaborativo para localizar artículos relevantes. Entre los proyectos que se tienen se puede mencionar:

GLOSS El proyecto del Servidor de Servidores de Glosario esta diseñado para localizar fuentes de información relevante.

SenseMaker herramienta que le ayuda a los usuarios a reformular iterativamente su información a través de organizar y activar grupos multidimensionales de búsqueda de resultados.

GASSROOTS Groupware para hallar información, combinar correo electrónico, noticias y web en un sencillo ambiente con distribución de

listas.

Un conjunto de estándares informales que negocian entre los vendedores de búsqueda y los usuarios para facilitar interoperaciones.

ComMentor Es una plataforma de anotaciones en el web que proporciona maneras de compartir información, y mantener notas.

InterOp Protocol. Es el corazón de "InfoBus", este protocolo describe métodos de acceso para buscar colecciones, adquirir resultados y encontrar fuentes.

InterBib es una herramienta para mantener información bibliográfica. Capaz de leer y escribir en muchos diferentes formatos, actúa como un repositorio unificado de registros bibliográficos.

La cantidad y diversidad de los proyectos mencionados, es un indicador del auge que está tomando el área de bibliotecas digitales. Como se ha visto es un campo multidisciplinario que cubre aspectos de áreas como recuperación de información, Interacción Humano-Computadora, protocolos de comunicación e Inteligencia Artificial. En la siguiente sección se presentan trabajos que involucran concretamente la tecnología de agentes.

3.5 AGENTES Y BIBLIOTECAS DIGITALES

3.5.1 Carnot

El proyecto Carnot está enfocado al problema de unificar información heterogénea y distribuida geográficamente. Proporciona al usuario medios para navegar a través de la información, para actualizar esa información consistentemente y para escribir aplicaciones sencillas para sistemas de información distribuidos. Carnot también incluye tecnología para visualización 3D de grandes espacios de información [Woelk 1994].

El sistema Carnot está implementado como un conjunto de agentes inteligentes. Para la integración de bases de datos heterogéneas, esos agentes son programados con el conocimiento necesario para buscar entre diferentes modelos de datos, lenguajes de consulta y esquemas de bases de datos. Este conocimiento es creado como el resultado de una fase de integración en la cual una nueva base de datos es descrita para Carnot en términos de su relación con los conceptos en un modelo libre. Cuando se recibe un "query", un agente es despachado al sitio remoto donde reside una base de datos. El agente recupera el dato requerido desde la base de datos remota usando el lenguaje de consulta apropiado, transforma los datos como sea necesario y regresa los datos formateados adecuadamente. Los agentes Carnot han sido desarrollados para varios

DBMSs (Ingres, Oracle, Sybase) y varios APIs y protocolos estándar (SQL Access Group Call Level Interface, ISODE Remote Database Access protocol). También se desarrollaron agentes para acceder a bases de datos textuales. Estos agentes soportan protocolos tales como Z39.50 y WAIS para la recuperación de términos de texto basados en palabras clave. Los agentes Carnot están implementados en un lenguaje llamado Rosette, es un lenguaje de comandos orientado a objetos, basado en el modelo de actores que incorpora soporte para acceso a procedimientos generados por otros lenguajes y ligado dinámico.

3.5.2 Modelo abierto de hipermedios basado en agentes

Este modelo se base en el uso de ligas dinámicas, conocidas como ligas de recuperación distribuidas [Salampasis et al. 1995], se desarrolló en la Universidad de Sunderland. Es un modelo que introduce la noción de agente de hipermedios. Los agentes de hipermedios son datos atómicos activos (por ejemplo, texto, imagen, video) conteniendo uno más componentes hipermedios los cuales son capaces de participar y operar cooperativamente en un ambiente de información distribuida. Cada agente de hipermedios esta dividido en dos partes: una cabeza del agente y el cuerpo del agente. La cabeza del agente contiene toda la información requerida para habilitar al agente La arquitectura de la cabeza del agente es independiente de la información fuente y su contenido. Diferentes fuentes de información (es decir, bibliotecas) son modelados como agentes hipermedios de alto nivel, llamados agentes de biblioteca. Estos agentes, deben ser capaces de interoperar en un ambiente distribuido, cooperativo y potencialmente heterogéneo. Se tienen ligas cooperativas, las cuales son ligas dinámicas diseñadas para proporcionar mecanismos de recuperación basados en búsquedas cooperativas para bibliotecas digitales distribuidas. Los usuarios pueden interactuar con su agente de hipermedios, crear, editar o refinar una liga cooperativa y finalmente activarse para iniciar un proceso de recuperación de información la cual esta basada en la cooperación entre diferentes agentes hipermedios (por ejemplo, agentes de biblioteca). Los agentes de bibliotecas reciben especificaciones para la búsqueda de documentos, pueden aplicar técnicas convencionales de recuperación de información para recuperar los documentos que coinciden con las especificaciones dadas, O pueden alternativamente comunicar las especificaciones de búsqueda a sus agentes de biblioteca cooperativos [Salampasis et al. 1995]. Este modelo usa ACL, un lenguaje de comunicación para agentes. ACL consiste de 3 partes su vocabulario, un lenguaje interno llamado KIF (Knowledge Interchange Format), y un lenguaje externo llamado KQML (**Knowledge Query and Manipulation Language**) [Finin et al. 1994].

3.5.3 AGS

En [Sanchez 1996; Sanchez y Leggett 1997] se describe AGS, una arquitectura abierta de bibliotecas digitales diseñada para proporcionar agentes como uno de los servicios disponibles a los usuarios. AGS propone manejar ciertas tareas en la biblioteca digital como procesos autónomos que pueden ser personalizados y controlados por el usuario final. Con AGS el usuario puede delegar tareas y continuamente interactuar con agentes en múltiples bibliotecas digitales vía una interfaz uniforme. AGS incluye en su diseño un *"toolkit"* especializado para facilitar la construcción de nuevos agentes. La implementación de AGS está basada en las tecnologías existentes cliente/servidor, sistemas de manejo de bases de datos y técnicas para modelar el comportamiento del agente.

AGS introduce cinco abstracciones en su arquitectura base: clases de agentes, instancias de agentes, acciones de agentes, perfiles de usuario y agencias. *Las clases de agentes* definen el comportamiento para las clases de usuario. Una clase de agente está definida por diversos atributos como pueden ser: el nombre de la clase, una descripción del comportamiento general del agente y requerimientos de sistema operativo para la operación del agente. *Instancias de agentes*. Una instancia de agente se asocia con una clase de agente cada vez que el usuario delega una tarea específica a un agente dado. *Las acciones* son objetos usados para construir bloques en la construcción de la funcionalidad del agente. Un *perfil de usuario* está asociado con cada usuario de AGS y contiene atributos arbitrarios para describir los intereses y preferencias del usuario. *Las agencias* se refieren a otras bibliotecas digitales conocidas por un servidor AGS dado. En AGS cada una de estas abstracciones pueden ser manejadas como objetos con atributos específicos que son usados por los subsistemas de AGS para facilitar el asignamiento y control de tareas para los agentes de usuario en la biblioteca digital.

Componentes de la arquitectura AGS. Por el lado del servidor se tiene el módulo *Active Library Service (ALiS)*, el cual proporciona funcionalidad para manejar las cinco abstracciones ya mencionadas. Los otros dos componentes que actúan directamente con ALiS son el *User Agent Manager (UAM)*, el cual proporciona la funcionalidad para la administración de los recursos de AGS, y el *User Agent Director (UAD)*, a través del cual el usuario final define y controla los agentes en la biblioteca digital. El orden para delegar una tarea a un agente en la biblioteca digital es el siguiente: el usuario selecciona una de las clases de agentes disponibles vía el UAD, el cual envía una instanciación a ALiS, ALiS recupera el código para implementar las acciones correspondientes a la clase de agente elegida y dependiendo de los atributos de la clase, permite al usuario proporcionar los parámetros requeridos y comenzar la ejecución en el servidor o en el cliente. Esta ejecución puede involucrar la creación de un nuevo proceso, la definición de reglas en la base de datos o el paso de código a un intérprete disponible. Una vez que la misión ha sido delegada, el usuario puede interactuar con sus agentes vía el UAD para suspender o reanudar sus tareas, inspeccionar el estado actual, o

eliminarlo. El UAD contacta a ALiS con la apropiada solicitud y ALiS envía un mensaje (sleep, awake, terminate) para las instancias involucradas y active o desactiva las reglas asociadas en la base de datos.

3.5.4 Agentes en FDL

En la sección 1.3.1 se describió el Proyecto de la Biblioteca Digital Florística (FDL), el cual pretende crear una biblioteca digital florística distribuida con información actualizada sobre todas las especies de plantas encontradas en áreas como Norteamérica, Mesoamérica y China.

Cuatro entidades primarias son mantenidas en una biblioteca digital florística: claves taxonómicas, mapas de distribución, ilustraciones y tratamientos. Las llaves taxonómicas son índices jerárquicos los cuales son usados por los botánicos para identificar especímenes basados en sus características. Las claves existen para cada nivel taxonómico (por ejemplo familia, tribu, género, especie, etc). Los mapas de distribución describen geográficamente el área geográfica exacta donde pueden hallarse especies particulares, mientras que las ilustraciones son interpretaciones artísticas de especímenes típicos en un grupo taxonómico dado. Los tratamientos *son el corazón de una flora*, ya que proporcionan descripciones morfológicas detalladas de plantas hasta nivel especie y niveles inferiores. Los tratamientos también se refieren a publicaciones que documentan las debilidades o naturaleza tóxica de las plantas, su estado con respecto al peligro de extinción, o discusiones sobre sus clasificaciones y nombres científicos. Cada descripción morfológica consiste de una serie de descriptores llamados estructuras, las cuales son abstracciones intermedias que agrupan un número (a veces implícito) de características de plantas para las cuales se puede asociar un valor [Sanchez et al. 1997, 1998]. Los tratamientos se enfocan a especies con un solo género, y cada uno es preparado y revisado por especialistas. Para crear un tratamiento, los investigadores estudian las plantas en el campo, examinan los especímenes en el herbario y revisan reportes publicados sobre trabajos previos. Existen aproximadamente 100 eventos discretos asociados con la edición y revisión de tratamientos, la mayoría de las actividades requieren coordinación a través de grandes distancias geográficas. Finalizado un tratamiento, se convierte de un documento en papel a un dato en línea en un sistema de manejo de bases de datos relacional con el fin de accederlo a través de varias interfaces y para extraerlo en varios formatos de publicaciones electrónicas y en papel. Aproximadamente 300 tratamientos conforman un volumen impreso solo para Norteamérica y se pretende crear 28 volúmenes. Esto significa que FDL debe manejar efectivamente un vasto número de relaciones intra-documentos, inter-documentos/intra volumene, e inter-volúmenes.

Un proyecto de estas dimensiones enfrenta grandes retos, entre ellos:

1. La introducción masiva de datos.
2. La búsqueda y recuperación de información relevante. En este sentido se han creado agentes que asisten al usuario tanto con la introducción de datos como en la recuperación de información.

3.5.4.1 *Chrysalis*

Este es un ambiente basado en agentes que facilita la tarea de introducir nuevos datos en la biblioteca digital florística FDL. Chrysalis, es un agente asistido para facilitar la generación de tratamientos. Cuenta con una herramienta para la edición de tratamiento que incluye un editor de tratamientos que incorpora automáticamente todas las características que aparecen en una llave taxonómica en la descripción morfológica apropiada para un tratamiento dado.

Cuando el autor introduce una descripción morfológica en el repositorio FDL, el agente compara la actual descripción con los tratamientos existentes en la biblioteca. Basado en una medida de similitud ajustable por el usuario, el agente determina si un tratamiento existente podría ser recuperado desde la biblioteca para que el autor puede usarlo como las bases para la descripción morfológica que está siendo generada. Si el usuario instruye al agente, las características del tratamiento sugerido son incorporadas, en el nuevo tratamiento y el autor puede añadir o remover características, ahorrando tiempo y asegurando consistencia.

Se usaron formas HTML para la introducción de estructuras, características y valores. Estas formas se invocaron desde programas hechos en C usando el protocolo vía C API y el "Web DataBlade" proporcionado por Illustra, la base de datos relacional orientada a **objetos** usada para mantener el repositorio de FDL. La interfaz del agente fue implementada usando JavaScript [Lopez 1997; Sanchez et al. 1997, 1998].

3.5.4.2 *FRA*

FRA es un agente que se incorporó a FDL con el objetivo de mejorar la búsqueda y recuperación de información relevante a los intereses y preferencias del usuario. FRA filtra y elimina aquella información que no considera de interés para el usuario. Este agente tiene la capacidad de trabajar de manera semi-autónoma porque el usuario en cualquier momento puede desactivarlo o suspenderlo, el usuario tiene la capacidad de calificar los documentos recuperados mejorando con esto próximas recuperaciones. El agente cuenta con el servicio de notificación vía correo electrónico para cuando nuevos resultados hayan sido obtenidos.

El esquema general del agente FRA esta compuesto por 6 módulos, que se interrelacionan entre sí enviándose mensajes [Cabrera 1997]. Para implementar el agente FRA, se tomó en consideración el ambiente distribuido en el que va a estar inmerso y el acceso a una base de datos

vía remota. El agente se desarrollo en Java, accedando una base de datos manejada por Illustra. Para lograr el acceso a la base de datos desde Java se usó el controlador *mijdbc*.

3.5.4.3 *MUTANT*

La biblioteca FDL ofrece a los usuarios diversas alternativas taxonómicas (puntos de vista de los taxonomistas), lo cual representa múltiples caminos a seguir en el sistema de hipermedios y por consiguiente una posible desorientación del usuario. Es en este contexto donde se desarrollo *MUTANT*, un agente que guía la búsqueda de información de acuerdo a las necesidades y objetivos del usuario y considera las diferentes opiniones o vistas taxonómicas [Flores 1997]. El objetivo de este proyecto es el de proveer el servicio de guías que faciliten la búsqueda y localización de información . Para el diseño y la implementación de *MUTANT* se utilizó Netscape Navigator, lenguaje de programación C, lenguaje de programación Java, manejador de Bases de Datos (DBMS) Illustra y un lenguaje de consulta de bases de datos (DM1L) SQL.

3.5.5 *Agentes en la Biblioteca Digital de Stanford*

El paradigma de agente ha sido usado en el Proyecto de la Biblioteca Digital de Stanford para hallar información (filtrado colaborativo) y para cuestiones económicas (tales como pagos).

3.5.5.1 *FAB: Un Agente Adaptativo para Automatizar búsquedas en el Web*

Es un sistema que aprende a navegar en Internet en nombre del usuario. Cada día le presenta una selección de páginas interesantes en el Web, el usuario evalúa cada página y da esta retroalimentación para que el sistema se adapte e intente producir mejores páginas al siguiente día. Se propone un sistema que ayude a los usuarios a mantenerse al día sobre los documentos nuevos y de importancia para el usuario que aparezcan en Internet . Más que soportar tareas de búsqueda, soporta tareas de navegación. Un aspecto definitivo de la actividad de navegar es que los usuarios no tienen una meta específica en mente.

Este sistema presenta los usuarios una selección de documentos que se piensa serian interesantes para ellos. Los usuarios entonces evalúan cada documento y el sistema ajusta sus parámetros para tratar de mejorar su desempeño. Los documentos son hallados usando búsquedas heurísticas en el Web, con los cambios heurísticos como la retroalimentación recibida del usuario. Esto es una tarea para una máquina de aprendizaje, en ninguna etapa los usuarios especifican un consulta en particular, pero si dan retroalimentación [Balabanovic et al. 1995]. Este sistema corre en ciclos discretos. El comportamiento de un ciclo se puede resumir como sigue:

1. Buscar en el Web, usando alguna heurística de búsqueda, tomando un límite de cantidad de tiempo.
2. Seleccionar las mejores páginas para presentárselas al usuario, usando alguna selección heurística.
3. Recibir una evaluación del usuario para cada página presentada.
4. Actualizar las heurísticas de selección y búsqueda de acuerdo a esta retroalimentación.

El sistema ha sido escrito en una mezcla de Python y C++. Para la recuperación de información se use una modificación del método de espacios vectoriales. La manera de acceder al sistema es a través de formas HTML.

3.5.5.2 *InterPay*

InterPay [Cousins et al. 1995] es una arquitectura para manejo de interacciones financieras en servicios de bibliotecas digitales. Esta diseñado para permitir interoperación entre diferentes servicios de pago, InterPay proporciona niveles de abstracción para permitir que las aplicaciones sean independientes de los detalles de mecanismos específicos de pago. El modelo contiene múltiples mecanismos de pago, modelos de interacción y políticas de cobro. Las componentes claves del modelo son agentes de pago y capacidades de pago que encapsulan políticas de pago. La arquitectura InterPay separa la tarea de acceso y pago de la información en tres capas: servicios, políticas de pago y mecanismos de pago. Esas capas corresponden a como los pagos ocurren en el mundo real. Para implementar InterPay se usó una tecnología de objetos distribuidos, lo que permite la construcción de entidades que directamente corresponden a los componentes en el modelo InterPay. En la capa de política de pago, un agente de pago negocia un modo aceptable de pago con el agente de colección. Los agentes de pago reciben y procesan facturas, validan el cargo a ser pagado e invocan una forma particular de pago que completará la transacción. Un agente de cobro es el que interactúa directamente con el agente de pago. Su responsabilidad es generar una factura de pago, y asegurar que el pago de la factura se haga en una forma apropiada. El agente de cobro es invocado siempre que el servicio desee imponer un cargo contra el usuario.

Cada medio disponible de pago (tales como tarjeta Visa, cuenta First Virtual, o dólares DigiCash) es encapsulado en una capacidad de pago. Así como los agentes de pago usan las capacidades de pago para acompañar las actividades específicas de pago, los agentes de cobro emplean capacidades de cobro para un mecanismo de pago en particular.

Se ha descrito una arquitectura que permite que varios mecanismos de pago puedan coexistir, soporta varias políticas de pago y simplifica las

interacciones del usuario en una biblioteca digital. Se han encapsulado los aspectos económicos de la biblioteca en un agente de pago y un agente de cobro, así que el usuario y el servicio están separados de los detalles específicos de mecanismos de pagos heterogéneos. Los detalles específicos de mecanismos de pago han sido también encapsulados en objetos, así que los agentes de pago y de cobro pueden ser programados por la gente que entiende de políticas de pago, pero no necesariamente de detalles de pago. Las funciones de pago están separadas de los servicios de la biblioteca.

3.5.6 Agentes en la Biblioteca Digital de Michigan

UMDL, la Biblioteca Digital de la Universidad de Michigan, pretende introducir una arquitectura abierta y distribuida donde agentes de software interactúen, cooperen y compitan para proporcionar los servicios de biblioteca [Birmingham 1995]. La investigación está enfocada a la construcción de un sistema que consiste de un conjunto cooperativo de 3 tipos de agentes de software. Los UIAs (agentes de interfaz de usuarios) manejan la interfaz con los usuarios. Expresan las consultas del usuario en una forma apropiada e interpretable por los agentes de búsqueda, mantienen perfiles de usuario basados en inferencias sobre los intereses del usuario, personalizando la presentación de los resultados de la consulta, y maneja los recursos disponibles del usuario con respecto a actividades de cuota por servicio, notifica al usuario de elementos de potencial relevancia para sus necesidades. Los Agentes Mediadores, proporcionan servicios de información intermedios. Ejecutan una variedad de funciones, por ejemplo, dirigen una consulta desde un UIA a una colección específica, monitorean el progreso de las consultas, transmiten resultados, traducen formatos, etc.

Existe una clase especial de agentes mediadores llamados facilitadores, para establecer equipos de agentes que ejecutan una tarea en común.

Los CIAs (Agentes de Colección de Interfaz) manejan la interfaz de UMDL con las colecciones, definen cuerpos del contenido de la biblioteca. Entre otras tareas, el CIA publica los contenidos y capacidades de una colección en un registro. Están asociados con cada colección específica y pueden manejar búsquedas en colecciones específicas de texto, imágenes, gráficas, audio y video. El sistema proporciona mecanismos para proteger acceso a la información y soporte operaciones de remuneración

Niveles de comunicación. Los agentes de UMDL se comunican en tres distintos niveles de abstracción. En el nivel más bajo, el agente emplea protocolos de red tal como TCP/IP para transportar mensajes. Los mensajes son interpretados y procesados por los protocolos de tarea específica. Inicialmente se definen protocolos especiales para la formación de equipos de agentes y la negociación de tareas, compartidas

por agentes en UMDL. Los protocolos UMDL representan el tercer nivel de abstracción en la comunicación de agentes.

El lenguaje que se usa es Conspectus Language (CL) con este lenguaje se describen las habilidades que son requeridas para participar en un equipo. Los tipos de mensajes UMDL son equivalentes a las "performativas" de KQML.

Se emplea un agente de registro, para mantener una base de datos que contiene información acerca de todos los agentes en el sistema UMDL, incluyendo el respectivo contenido y descripción. El agente de registro acepta expresiones de descripción de CL y actualiza la base de datos de acuerdo a esto. Además del agente de registro, se han implementado varios tipos de agentes mediadores, el planeador de query, el facilitador de mercado y remora.

También se crearon los planeadores de tareas (TPA; Task Planning Agent): estos son una clase de agentes que realizan tareas de búsqueda. Los mediadores de planeación de consultas son una subclase de agentes de planeación de tareas que localizan en las colección información que satisfaga el query del usuario.

Son responsables de descomponer las tareas complejas en otras más simples y de formar equipos de agentes que ejecuten dichas tareas. El TPA se comunica asíncronamente con una capa de comunicación la cual es responsable de manejar los mensajes a bajo nivel [Vidal y Durfee 1996].

Además de los planeadores de tareas, se creó un agente mediador llamado *remora*, que ofrece servicios de notificación para una variedad de recursos de biblioteca. A través de este agente, los usuarios pueden especificar eventos de interés, tales como nuevos elementos que coinciden con las descripciones que aparecen en las colecciones, y recibir notificaciones cuando tal evento ocurra. Este agente trabaja en nombre de los agentes de usuario UIAs aceptando scripts que especifican eventos de interés y acciones a ejecutar cuando los eventos ocurran [Atkins et al. 1995].

Finalmente, en la última sección de este capítulo, se van a presentar algunos de los métodos que existen para recuperación de información, estos métodos son de gran utilidad al ser incorporados en bibliotecas digitales, ya que facilitan la búsqueda, recuperación y filtrado de datos.

3.6 METODOS DE RECUPERACIÓN DE INFORMACIÓN

(3.2b)

En un modelo booleano extendido, los documentos conteniendo

solamente un término de una consulta *and* pueden ser recuperados con un valor de recuperación bajo. Los documentos conteniendo varios términos de una consulta o son preferidos sobre aquellos conteniendo solamente un término.

Las desventajas del modelo extendido son:

1. La mayoría de las leyes de procesamiento básico booleano, tales como leyes de asociación y distribución, no son estrictamente válidas en el sistema extendido.
2. La aplicación recursiva de las fórmulas (3.2) producen complicadas expresiones para declaraciones booleanas con un gran número de operadores.

Como ejemplos de sistemas que usan estos modelos de recuperación se pueden mencionar los siguientes:

El Sistema STAIRS, es un sistema de recuperación de información de IBM. STAIRS un acrónimo para "Storage And Information Retrieval System", un sistema que proporciona facilidades para recuperación de texto tanto de palabras como de frases hechas a partir de combinaciones booleanas complejas STAIRS use el modelo de recuperación booleano convencional [Blair y Maron 1985].

El sistema Smart (Sistema para la Manipulación y Recuperación de Textos), fue diseñado en los 60's como un sistema prototipo para recuperación y procesamiento de texto. El sistema Smart esta basado en el modelo de espacios vectoriales. Este sistema usa técnicas de indexado automático para generar los vectores de las consultas y los vectores documentos a partir de consultas y documentos disponibles [Salton y Buckley 1988].

$$\sqrt{(d_A - 0)^2 + (d_B - 0)^2}$$

$$\sqrt{(1 - d_A)^2 + (1 - d_B)^2}$$

$$\text{sim}(D, Q_{(A \text{ or } B)}) = \sqrt{\frac{d_A^2 + d_B^2}{2}}$$

$$\text{sim}(D, Q_{(A \text{ or } B)}) = \sqrt{\frac{(1 - d_A)^2 + (1 - d_B)^2}{2}}$$

Tabla 3.2 Valores de Similitud Query-documento para Sistemas Convencionales y Extendidos(adaptada de [Salton et al. 1983]) a) Recuperación Booleana Convencional

	Términos		Similitud con el query	
	A	B	(A or B)	(A and B)
Documento D1	1	1	1	1
Documento D2	1	0	1	0
Documento D3	0	1	1	0
Documento D4	0	0	0	0

Tabla 3.3 Valores de Similitud Query-documento para Sistemas Convencionales y Extendidos(adaptada de [Salton et al. 1983]) b) Recuperación extendida

	A	B	(A or B)	(A and B)
Documento D1	1	1	1	1
Documento D2	1	0	1/raiz de 2	1-(1/raiz de 2)

Documento D3	0	1	1/raiz de 2	1-(1/raiz de 2)
Documento D4	0	0	0	0

$$1/\sqrt{2}$$

$$1 - 1/\sqrt{2}$$

$$\text{sim}(D, Q_{(A, a \text{ or } B, b)}) = \sqrt{\frac{a^2 d_A^2 + b^2 d_B^2}{a^2 + b^2}}$$

$$\text{sim}(D, Q_{(A, a \text{ and } B, b)}) = 1 - \sqrt{\frac{a^2 (1 - d_A)^2 + b^2 (1 - d_B)^2}{a^2 + b^2}}$$

3.7 RESUMEN

Por lo que se ha visto a lo largo de este capítulo se puede concluir que la tecnología de agentes es una excelente herramienta para la búsqueda, filtrado y recuperación de información, dentro de los ambientes distribuidos, heterogéneos y multimedios que engloban a las bibliotecas digitales.

En el siguiente capítulo se presenta el diseño de un agente móvil recuperador de información en bibliotecas digitales florísticas. Esta es una aplicación real que engloba a las tres áreas mencionadas en los capítulos 2 y 3 - *agentes, bibliotecas digitales y técnicas de recuperación de información*.

Pérez Lezama, C. V. 1998. **Agentes Móviles en Bibliotecas Digitales**. Tesis Maestría. Ciencias con Especialidad en Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas Puebla. Mayo. Derechos Reservados © 1998.