VNIVERSIDAD NACIONAL
AVFNPMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA
CIUDAD UNIVERSITARIA

DISEÑO E IMPLANTACIÓN DE UN BANCO TERMINOLÓGICO PARA EL GRUPO DE INGENIERÍA LINGÜÍSTICA

T E S I S

QUE PARA OBTENER EL GRADO DE:

INGENIERO EN COMPUTACIÓN

P R E S E N T A:

GABRIEL GARDUÑO TORRES

DIRECTOR: DR. GERARDO SIERRA MARTÍNEZ

MÉXICO, D.F. 2003

TESIS
FALLA
DE
ORIGEN

Para mi amada esposa
con la que siempre conté con
su apoyo incondicional
y amor en todas las facetas
de mi carrera profesional

Para mis padres
que siempre tuvieron
plena confianza en mi
y fueron mi puntal durante
muchos años

A toda mis hermanos,
familia y amigos
que pusieron un granito
para que se lograra este trabajo
y dejaron alguna huella en mi vida

Agradecimientos

Un agradecimiento especial al Dr. Gerardo Sierra por apoyarme en el desarrollo del presente trabajo, en una parte de mi formación académica y darme la oportunidad de abrirme un panorama más amplio en mi futuro profesional.

A todos los integrantes del Grupo de Ingeniería Lingüística que enriquecieron esta tesis con sus comentarios y seminarios.

A la Facultad de Ingeniería que me dio la oportunidad de formarme profesionalmente.

Al Instituto de Ingeniería y CONACyT que me albergaron como becario y me patrocinaron durante todo el proceso de este trabajo.

A mi Alma Mater la Universidad Nacional Autónoma de México la cual ha sido mi lugar de enseñanza durante muchos años.

Índice

Índice de figuras

Índice de tablas

Introducción

Antecedentes

En la época actual existe la tendencia de apoyarse en la inteligencia artificial para elaborar sistemas expertos inteligentes, pero ¿qué pasa con las bases de conocimiento de tales sistemas? ¿dónde residen? ¿cómo se mantienen? La forma de darle una respuesta a estas preguntas es contar con un banco terminológico en el cual los desarrolladores tengan la posibilidad de manipular su información y contar con un lugar fijo en donde tenerla residente.

El auge de las enciclopedias electrónicas va en aumento día con día, pero no existirían si no se hubiera desarrollado la investigación en los bancos terminológicos. Las enciclopedias básicamente están compuestas de términos con sus definiciones respectivas, aunque en este caso las definiciones son enciclopédicas, nos brindan la oportunidad de hacer búsquedas entre los términos relacionados y los sinónimos pertenecientes a un término en específico. Además cuentan con bancos de imágenes y videos que están interrelacionados con el banco terminológico de la enciclopedia.

Con la aparición de los bancos terminológicos, la carga de los expertos en traducción se ha disminuido considerablemente, ya que anteriormente tenían que estar consultando cantidades de libros con las diferentes equivalencias en otros idiomas del término en cuestión. Sin embargo, ahora tienen la disponibilidad de contar con los bancos en formato electrónico, los pueden consultar vía Internet y disponen de una gran cantidad de información almacenada en tan solo uno de ellos.

El trabajo que se presenta a continuación es el resultado de la investigación realizada durante la estancia como becario del Grupo de Ingeniería Lingüística (GIL), del Instituto de Ingeniería, UNAM. El proyecto principal que tiene el GIL es la elaboración de un diccionario integral con capacidad de hacer búsquedas onomasiológicas, es decir, encontrar el término a partir del concepto introducido por el usuario. De este proyecto se desprenden distintas líneas de investigación y una de ellas pertenece al "Diseño e implantación de un banco terminológico para el Grupo de Ingeniería Lingüística".

Objetivo de la tesis

El objetivo de la presente tesis es elaborar un banco terminológico con los siguientes fines:

- Elaborar el banco como integrante de la base de conocimiento léxico la cual es parte integral del diccionario onomasiológico.
- Recopilar toda la información terminológica que el GIL ha extraído de los textos de especialidad y que se debe encontrar almacenada en un mismo sitio.

- Contar con una herramienta implantada en Internet para que apoye a otras líneas de investigación que se están elaborando dentro del GIL y que los usuarios de ésta herramienta tengan acceso a ella desde cualquier lugar.

Metodología

La programación de software necesita apoyarse en principios sólidos y firmes que faciliten el desarrollo de la actividad de programación, estos principios los facilita la ingeniería del software.

Un modelo estándar para el desarrollo de un buen sistema es conocido como: "El ciclo de vida del software[1]". El modelo del ciclo de vida del software brinda un panorama de las actividades que ocurren durante el desarrollo del sistema, el cual pretende determinar el orden de las etapas involucradas y los criterios de transición asociados entre estas etapas.

Para llevar a cabo la construcción de una aplicación informática. antes de plantearnos cómo vamos a hacer las cosas nos tendremos que plantear el problema en sí y qué es lo que hay que hacer, es decir, como en cualquier otra rama de la ingeniería, tenemos que pasar por ciertos pasos para resolver el problema en cuestión.

- Formular el problema.
- Plantearse objetivos.
- Buscar y desarrollar soluciones.
- Evaluar las distintas soluciones en función de los objetivos.
- Refinar y verificar la solución escogida.

Algo que nos es muy útil en el desarrollo de un sistema es que el ciclo de vida se apoya en la teoría "divide y vencerás"[2]. Esta es una técnica de diseño de algoritmos que consiste en resolver un problema lo suficientemente grande, subdividiéndolo en problemas más pequeños. Si los problemas ya divididos son todavía relativamente grandes se aplicará de nuevo la técnica hasta alcanzar problemas lo suficientemente pequeños para ser solucionados directamente.

La planificación del ciclo de vida es un factor principal para conseguir los objetivos buscados. En el desarrollo de la presente tesis se decidió planificar con base en el modelo de "Ciclo de vida en cascada pura", es decir, el desarrollo del sistema va progresando a través de una secuencia ordenada de etapas generales tales como:

[1] http://web.madritel.es/personales3/edcollado/ingsw/tema2.htm
[2] http://www.lcc.uma.es/~av/Libro/CAP3.pdf

- *Definición de requisitos*: esta etapa se centra en la formulación correcta del problema en cuestión y los requerimientos que se necesitan para resolverlo.

- *Análisis*: en esta etapa hay que expresar la estructura de la solución, centrándonos en las necesidades del usuario, de tal forma que éste pueda comprender sus necesidades de la manera más correcta posible. En esta etapa nos aseguramos de haber entendido las necesidades del usuario.

- *Diseño*: en esta etapa se ha de expresar el problema y la solución en términos informáticos para que el programador pueda realizar su trabajo.

- *Codificación*: en esta etapa se generan los programas y las bases de datos que componen la aplicación.

- *Pruebas*: en esta etapa se comprueba que, en su conjunto, todos los componentes de la aplicación funcionan correctamente.

- *Mantenimiento*: con la utilización de la aplicación se van descubriendo funcionamientos defectuosos y/o ausencia de funcionalidad que no habían sido detectados con anterioridad. Como el código libre de errores al 100% es prácticamente imposible de conseguir, hay que proceder a la corrección de errores y comenzar nuevamente desde el principio con todas las etapas del ciclo de vida.

El sistema se va revisando tras cada una de las etapas y se tienen que haber conseguido todos los objetivos de la etapa anterior para poder pasar a la siguiente, es un proceso secuencial.

Existen una gran cantidad de metodologías a seguir para construir un sistema. En el desarrollo de la presente investigación nos basamos en la "metodología esctructurada" elaborada por Yourdon, Marco, Winberg y Jacson en los 80's[3]. Esta metodología es un conjunto de técnicas que sirven para hacer un modelo del problema de tal forma que pueda evolucionar gradualmente hasta obtener el modelo de la solución final.

En este marco, el desarrollo del banco terminológico está basado en el ciclo de vida en cascada pura, aplicando la técnica de divide y vencerás, utilizando la metodología estructurada. Todo lo mencionado anteriormente está aplicado en cada una de las etapas del desarrollo del sistema y se detalla en los capítulos aquí presentados.

Resumen de los capítulos

En el capítulo 1 hacemos un análisis exhaustivo de los orígenes de los bancos terminológicos, a quiénes ayudan, en qué áreas son funcionales, así como su evolución, de qué manera se clasifican, qué es lo que debe contener en esencia un banco y, para finalizar el capítulo, se hace un análisis de distintos bancos que se encuentran en Internet.

[3] http://www.bibliodgsca.unam.mx/manuales/manual.pdf

En el capítulo 2 consideramos los requerimientos del sistema basándonos en una serie de preguntas frecuentes (FAQ. Frequently Asked Questions) desarrollada por Ventura Miranda en su tesis de licenciatura. Las preguntas están elaboradas de manera que nos van introduciendo en la problemática de lo general a lo particular. Las preguntas se tienen clasificadas en 7 series: contexto, problemática, volumen y tipo de información, perfil y participación de los usuarios, características deseadas, infraestructura y tipología.

En el capítulo 3 se menciona todo el preámbulo que se tuvo para obtener como resultado las bases de datos. También mostramos el análisis y diseño del sistema, basándonos en el ciclo de vida del software.

En el capítulo 4 se muestra la integración de las bases de datos con el software previamente programado para llegar a la implantación del banco terminológico y se presentan las diferentes interfaces con que cuenta el sistema.

En el capítulo 5 consideramos al banco terminológico como herramienta que apoya a diferentes líneas de trabajo que se están llevando a cabo en el GIL.

Finalmente se presenta un capítulo con las conclusiones que se obtuvieron al desarrollar la presente tesis.

Capítulo 1. Bancos terminológicos

Una base de datos es un sistema formado por un conjunto de datos con su respectivo software para gestión de los mismos. El software encargado de la gestión permite:

- Controlar el almacenamiento de datos redundantes.
- Independencia entre los datos y los programas que los usan.
- Almacenamiento conjunto de los datos y sus relaciones entre ellos mismos.
- Acceso a los datos de distintas formas.

En una base de datos se almacena información de una serie de objetos o elementos. Estos objetos reciben el nombre de entidades. Una entidad es cualquier cosa sobre la que se almacena información.

De cada entidad se almacenan una serie de datos que se denominan atributos de la entidad. Puede ser atributo de una entidad cualquier característica o propiedad de ésta.

En una base de datos la información de cada entidad se almacena en registros, y cada atributo en campos de dicho registro.

1.1 Definición de banco terminológico

Se define un banco de datos terminológico como una colección de bases de datos que contienen conjuntos de datos estructurados, fiables y homogéneos almacenados en una computadora. La información contenida en un banco es sobre una temática de carácter científica y/o tecnológica, y contiene las terminologías con sus respectivas definiciones catalogada en diferentes áreas temáticas.

Los bancos terminológicos tienen como misión fijar y difundir el vocabulario especializado de los ámbitos de la ciencia, tecnología y en general lo relativo a los campos especializados del saber y el conocimiento.

Para que los datos sean estructurados, fiables y homogéneos, deben de cumplir con ciertas propiedades:

- Ser compartidos por diferentes usuarios y/o aplicaciones.
- Permitir el acceso directo a ellos mismos.
- Contar con un conjunto de programas que los manipulen.

- Estar estructurados independientemente de las aplicaciones y del soporte de almacenamiento que los contiene.

- Presentar la menor redundancia posible.

La información que contienen los bancos se va recopilando de una gran variedad de fuentes (libros, revistas, enciclopedias, corpus de textos especializados, etc.). Regularmente está compuesta de términos especializados, definiciones, fuentes de donde fue extraída la información y datos que son necesarios para cumplir con los objetivos de cada uno de los desarrolladores de los bancos. Además, la información se encuentra automatizada, relacionada entre sí y estructurada con respecto a la importancia que se le vaya dando a cada campo.

Usualmente los bancos tienen una base de datos principal en donde se encuentran los términos almacenados. De igual modo poseen otras bases que están interrelacionadas entres sí, las cuales van a servir para proveer al usuario de información adicional de algún aspecto de los términos.

Hay una gran variedad de usuarios de los bancos terminológicos, éstos hacen uso de ellos para poder cumplir con sus necesidades específicas y darle respuesta a sus preguntas sobre cuestiones concretas, en un tiempo muy breve.

Se mencionan enseguida a los diferentes tipos de usuarios existentes[1]:

- Los profesionales de áreas especializadas: tienen diferentes necesidades de manipular la información con que cuentan. Sin embargo, la manera de manejar tal información no es nada fácil y en algunos casos no es posible tener acceso directo a la misma, por lo que los bancos son de una gran ayuda para ellos. Una gran ventaja para los especialistas radica en contar con una gran cantidad de información almacenada en un solo lugar, por lo que la cantidad de tiempo invertida en buscarla se reduce al mínimo. Dentro de estos profesionales encontramos a los:

  a) Especialistas de la comunicación, la información y la documentación.

  b) Expertos en lexicografía, terminología y traducción.

  c) Ingenieros lingüistas y lingüistas computacionales.

  d) Editores, profesores de lengua, investigadores en lingüística aplicada.

  e) Profesores de áreas especializadas.

- El público en general: toda aquella persona que tenga la posibilidad de tener acceso a Internet, tiene la facilidad de hacer búsquedas de información sobre

[1] Cabré (1993)

términos especializados con una temática en concreto, de una manera fácil, sin costo alguno y con una gran variedad de bibliografía en donde consulten la información, esto le ahorra tiempo dinero y esfuerzo al usuario.

La capacidad de almacenar grandes cantidades de términos con sus respectivas pesquisas, de mantener la información actualizada de forma más sencilla y con menor costo, y el que se pueda difundir a gran escala utilizando sistemas más actuales, han convertido a los bancos terminológicos en herramientas imprescindibles para todo tipo de usuarios.

1.1.1 Funcionalidad

Los bancos terminológicos tienen diferentes funciones, dependiendo del área a la que estén enfocados, como por ejemplo:

- La inteligencia artificial y la ingeniería del conocimiento precisan de los términos para poder realizar su trabajo.
- Las bases de conocimiento que contienen los sistemas expertos necesitan un banco terminológico donde residir físicamente.
- A los traductores se les facilita el trabajo proporcionándoles una herramienta de consulta única que comprenda los distintos diccionarios, cómodos para su consulta y con propuestas fiables.
- Los especialistas pueden disponer de una gran cantidad de información en una sola obra de referencia.
- Cualquier persona puede acceder a un diccionario de temática especializada, en la mayoría de los casos de forma gratuita, con solo contar con una conexión a Internet.

Nombrando algunas instituciones que hacen uso de los bancos terminológicos tenemos al Fondo Monetario Internacional, la Organización Internacional del Trabajo, por mencionar algunas. Para ahondar más en el tema se puede observar el apartado 1.5 de este capítulo.

1.2 Evolución

Anteriormente había ciertos trabajos que únicamente el terminólogo los podía llevar a cabo manualmente, esto resultaba extenuante y tardado. Con la aparición de los bancos terminológicos tuvieron una gran cantidad de beneficios ya que la computadora puede hacer de forma autónoma gran parte de dichas labores, o bien lo asiste para realizarlas de forma semiautomática. Dentro de estos trabajos encontramos:

- Vaciado automático y semiautomático de documentos.

- Redacción de fichas automatizadas y semiautomatizadas.
- Elaboración de términos y definiciones.
- Creación de bancos de datos a partir de textos.
- Verificación ortográfica y gramática de documentos.

Hay diversos factores que han provocado que la concepción de los bancos fuera cambiando con el tiempo, de los cuales destacan la necesidad de obtener y mantener grandes cantidades de información en un mismo lugar, el gran desarrollo que han tenido la computación y la tecnología de las computadoras, los avances que se han producido en la elaboración de sistemas y software, así como la aceptación y mayor uso de las computadoras por parte de los usuarios.

Conforme han ido evolucionando los bancos, han pasando por diferentes etapas, las cuales clasificamos a continuación en tres distintas generaciones.

1.2.1 Bancos de primera generación

Los primeros bancos de datos terminológicos se crearon a principios de la década de los setenta. En este tiempo el poco desarrollo de la computación no permitía elaborar sistemas eficaces que facilitaran el uso de los bancos, y de igual modo la falta de tecnología en las computadoras sólo permitía concebirlas como instrumento de almacenamiento de información, con sus limitantes. A causa de estos motivos, eran bancos de acceso restringido por la dificultad de búsqueda de información y normalmente los bancos precisaron de un intermediario para su uso.

Inicialmente estos bancos fueron concebidos como meros instrumentos al servicio de la traducción en la que únicamente se tenía información de palabras y su equivalencia en algún otro idioma; después se comenzaron a implantar dentro del ramo de la biblioteconomía, para finalmente aplicarlo a la terminología.

El factor prioritario de evaluación en esta clase de bancos era la cantidad de términos que contenían, el número de áreas temáticas que cubrían y la cantidad de información sobre cada término. Había cuestiones que eran problemáticas en esos tiempos:

- No obstante que las computadoras eran consideras como un lugar para almacenar información, la capacidad de ellas era muy limitada, por lo que tenía que haber una gran cantidad de dispositivos de almacenamiento secundario para el acopio de los datos.
- La compatibilidad de los sistemas así como la migración de información entre una computadora y otra.
- El tiempo que se invertía en hacer una consulta al banco y recuperar la información era muy elevado.

En determinado momento se encontró un problema bastante grave en este tipo de bancos: en un principio, la mayoría de ellos fueron desarrollados con sistemas creados para cubrir las necesidades que se tenían al momento, y el diseño al que fueron sometidos fue muy limitado. La consecuencia de esto fue que dichos bancos fueron incapaces de cubrir nuevas exigencias que se presentaron a futuro, por lo que se tuvo que pensar en rediseñar a la gran cantidad de bancos que tenían este problema.

Desde 1980 los especialistas en ingeniería lingüística y computación comenzaron a trabajar en la mejora de estos bancos, para poder obtener una mejor calidad y actualidad de los datos que contenían, así como la facilidad de acceso a la información. Sin embargo, se dieron cuenta que rediseñar completamente sus bancos tenía un costo muy elevado, por lo que optaron por volver a implantar sus bancos y empezar desde cero, lo cual dio origen a la siguiente generación.

1.2.2 Bancos de segunda generación

Esta generación comenzó dándole énfasis a la calidad de los datos, la selección de información sobre los términos en función de los usuarios, la especialización de los datos, el nivel de actualidad y novedad de los mismos, así como suprimir el problema de la compatibilidad entre las distintas computadoras.

En esta etapa se concibieron bancos más innovadores en lo que respecta a la arquitectura de las bases de datos así como a la lógica y a la estructura de la información contenida en ellas. Así pues, los bancos resultaron ser más flexibles y efectivos para servir a las necesidades de los usuarios.

El tratamiento más profundo de la información contenida en los bancos, así como la generación de herramientas (software) que facilitaran la comunicación persona-computadora, dio como resultado:

- Minimización del tiempo que se requería para una consulta y para la recuperación de información.
- Las formas de consulta fueron más variadas, eficaces y amigables para el usuario.
- Los usuarios realizaron sus consultas de manera autónoma, es decir, ya no tienen la necesidad de un intermediario para el uso del banco.
- En esta generación, las interfaces en lenguaje natural estaban en pleno auge de investigación y desarrollo.
- Implantación y desarrollo de muchos bancos de pequeñas dimensiones y de temática muy especializada.

- Los desarrolladores de los nuevos bancos no tuvieron el lastre de tener que transformar un sistema establecido anteriormente, como había sucedido con los de la primera generación.

- Fueron posibles las consultas remotas de información mediante una PC y un módem, esto es, los usuarios ya pueden acceder a los bancos terminológicos vía Internet.

- Se hicieron disponibles bancos terminológicos en CD-ROM.

Este tipo de bancos terminológicos se siguen elaborando aún en nuestros días, pero la gran importancia que tienen hoy la inteligencia artificial y los sistemas expertos, como apoyo para el desarrollo de los bancos, está dando como resultado una nueva generación de bancos.

1.2.3 Bancos de generación inteligente

En nuestros días la elaboración de los bancos terminológicos. tiende a ser de pequeñas dimensiones (especializados en cuanto a su área temática) e inteligentes, ya que pueden descifrar adecuada y selectivamente las necesidades de información tanto en lo que se refiere a los datos de búsqueda así como a las informaciones que les acompañen, de unos usuarios cada vez más diversificados, informados y exigentes.

En esta etapa forman parte los trabajos sobre sistemas, desarrollados en el marco de la inteligencia artificial. El problema de la migración de la información ya no causa ninguna dificultad ni aún en diferentes sistemas operativos. La computadora utiliza la información que contiene almacenada como elemento de partida para realizar. por su propia cuenta, otras operaciones más complejas y con mayor semejanza a las que haría un ser humano. El objetivo es que las interfaces de lenguaje restringido sean en lenguaje natural y que además el sistema sea capaz de poder emular el pensamiento humano, por ejemplo, que la computadora sea capaz de detectar errores al hacer consultas como "definición del término balín de forma cuadrada".

Este tipo de banco terminológico es el futuro en la investigación; se va a evolucionar de una manera muy significativa el concepto que se tiene de ellos hasta el momento, además de que van a ser de gran ayuda para todos los usuarios, fáciles de acceder, manejar, y lo mejor de todo es que con ayuda de la inteligencia artificial y los sistemas expertos, se podrán tener bancos que lleguen a pensar como el ser humano.

1.3 Clasificación

No hay una tipología de bancos de datos terminológicos capaz de avalar las múltiples facetas que éstos puedan tener, sin haber tomado varios criterios de categorización. Como consecuencia, no puede haber una clasificación única, sino que debe de ser múltiple, de acuerdo con los parámetros que tenga el banco para cada caso.

El siguiente ordenamiento que se exhibe, no implica que forzosamente los bancos deben de estar clasificados con todas y cada una de las partes que se mencionan, sino que pueden tener una o una combinación de varias de ellas, esto depende de los parámetros que se estén considerando en el banco.

En nuestro caso en particular, mencionamos los puntos que consideramos más importantes para nuestra investigación, tomando como referencia la clasificación de los bancos que presenta Cabré[2].

- Por sus objetivos:

  a) Bancos informativos: Difunden la terminología de uno o varios ámbitos.

  b) Bancos equivalentes: Facilitan las unidades equivalentes en otras lenguas.

  c) Bancos prescriptivos: Intervienen en el uso de los términos así como en su corrección.

  d) Bancos descriptivos: La información que contiene es tal y como la presenta el autor, sin hacerle modificación alguna.

  e) Bancos normativos: Para estar incluida dentro del banco, la información tiene que seguir una serie de normas. Las normas dependen de las personas asignadas para tal caso, dentro del desarrollo del banco.

- Por la organización de sus datos:

  a) Bancos organizados por el término: Cubren consultas a partir del término, para que nos guíe hasta el concepto.

  b) Bancos organizados por el concepto: Se parte de un concepto o una idea, para que con base en ella nos lleve hasta el término buscado.

- Por su temática:

  a) Bancos de temática especializada general: Contienen información sobre distintas áreas, con una gran cantidad de términos en campos diversos aunque no suelen tratar los campos con la misma profundidad.

  b) Bancos de términos de un solo dominio especializado: Son de dimensiones más reducidas que los anteriores, pero con capacidad de poder ser actualizados con mayor facilidad y menor costo, ya que únicamente tratan una sola área temática a la vez.

[2] Cabré (1993)

- Por el interés prioritario de los datos que contienen:

  a) Bancos de términos: Normalmente va el término junto con su definición.

  b) Bancos de términos contenidos en documentos: Los términos, sus respectivas definiciones e información adicional se encuentran inmersos en los textos especializados.

  c) Bancos enciclopédicos: Términos con definiciones obtenidas de enciclopedias.

  d) Bancos visuales: Términos que tienen su propia imagen la cual hace alusión al mismo.

- Por el número de idiomas de las informaciones terminológicas:

  a) Bancos monolingües: Están elaborados completamente en un sólo idioma.

  b) Bancos plurilingües: Los conceptos poseen una ficha completa de informaciones sobre el término en cada uno de los diferentes idiomas.

  c) Bancos monolingües con equivalencias en otros idiomas: La información completa del término se presenta sólo en un idioma y se asignan las equivalencias denominativas en los demás idiomas.

- Por el hardware del banco:

  a) Bancos desarrollados en macrocomputadoras. La información contenida está compuesta por grandes bancos terminológicos. por lo que se necesitan computadoras lo suficientemente grandes para poder tener residente la información.

  b) Bancos desarrollados en microcomputadoras. Son bancos desarrollados por centros especializados en ciertas áreas temáticas conteniendo información más actual e innovadora o por algún profesional que los elabora de acuerdo con sus necesidades y normalmente se encuentran residentes en una PC.

- Por el tipo de banco:

  a) Bancos de primera generación.

  b) Bancos de segunda generación.

  c) Bancos de generación inteligente.

1.4 Constitución

Hay que tomar en cuenta que todo banco terminológico, independientemente de las funciones o las dimensiones que vaya a tener, debe de contar con una entidad principal que contenga a los términos. Tanto la entidad de términos, como todas las demás entidades adicionales, están integradas por registros, cada uno de los cuales contiene todos los datos relacionados a un mismo término.

En la constitución de un banco terminológico, se tienen tres grandes etapas: la de recopilación de información o entrada (1.4.1), la de almacenamiento (1.4.2) y la de recuperación de información o salida (1.4.3).

La siguiente constitución de bancos terminológicos está basada conforme a nuestras necesidades, de la constitución que presenta Cabré[3].

1.4.1 Recopilación de información

En esta primera etapa de la constitución se tiene que considerar el material que se va utilizar como base, tomando en cuenta que se va a partir de textos y no de términos. La información puede provenir de los diccionarios o listas de términos procedentes de otros bancos o publicaciones, pero en última instancia conforman en su conjunto un texto, a partir del cual se va a realizar el vaciado de información. Por este es importante tomar en cuenta:

- Un plan previo totalmente detallado sobre el proceso de recopilación del material y estructura de los registros del banco. No hay que olvidar que la información se está actualizando de forma permanente, por consiguiente el proceso de estructuración del banco nunca termina.

- Las características con las que va a contar el material (área(s) temática(s) pertinente(s), la selección y evaluación de los textos a utilizar, datos que se deben de recopilar de cada texto).

- La localización, esto es, el lugar donde se encuentra el material o en qué tipo de soporte magnético está almacenado. Es importante que se tenga el material digitalizado, debido a que es el punto de partida de la automatización de todos los procesos.

- El método de extracción de la información, esto es, cómo se va a pasar de los textos a los términos y a sus datos adicionales. La extracción puede ser manualmente, identificándose los términos en los textos o de fichas previamente llenadas, para después introducirlos directamente en la base de datos. Si se tienen disponibles textos digitalizados, puede aplicarse un programa de vaciado automático, aunque en este último caso las herramientas disponibles para el

[3] Cabré (1993)

vaciado no se han perfeccionado en su totalidad, por lo que el proceso es semiautomático.

- La descripción de manera detallada acerca de la información que va a contener el banco sobre cada término (fuente, autor, equivalencias, etc.), ya que con esta información se va a llenar el registro de vaciado. Los diferentes datos contenidos en el banco deben formar parte de su campo correspondiente, y estar representados de forma estructurada (en algunos casos debe ir codificada) para facilitar posteriormente la recuperación de la información por cualquier campo posible.

1.4.2 Almacenamiento

El resultado que vamos a obtener después de la etapa anterior será una entidad de términos que es propiamente la base de datos más importante del banco terminológico. Sin embargo, los términos no componen todo el banco, se debe de complementar con otras bases que especifiquen datos que contengan los registros terminológicos, los cuales nos proporcionan otro tipo de información.

Enseguida se va a ver en detalle el tipo de información que debemos tener dentro del banco terminológico.

1.4.2.1 Los registros terminológicos

En esta sección se aborda la información que tiene que ver directamente con los términos.

- El *término*, es decir, la palabra o grupo de palabras introducidas por un nexo preposicional[4]. Viene siendo el registro principal y además constituye la entrada del banco terminológico.

- La fuente de donde proviene la información contenida en el banco. En algunos casos la fuente puede venir codificada, si este es el caso, se debe de tener una base de datos complementaria en la que va a contener una entidad con las fuentes de referencia. Se puede o no tener un enlace entre la entidad de fuentes y un banco de datos bibliográfico en donde se pueda hallar la descripción completa de cada referencia.

- Datos sobre cuestiones lingüísticas: formas abreviadas, sinónimos, las relaciones entre los mismos términos, etc.

- Contexto definitorio de los términos.

- Equivalencias en otros idiomas.

[4] Cerdá (1986)

- Datos sobre la gestión de la información: responsable del ingreso de la información al banco, fecha de registro, datos de los usuarios, etc.

1.4.2.2 Las entidades complementarias

Generalmente la estructura de un banco terminológico incluye las siguientes entidades:

- Entidad de términos (entidad principal que no debe de faltar en un banco terminológico).
- Entidad de fuentes.
- Entidad de áreas temáticas.
- Entidad de responsables.
- Entidad de equivalencias.
- Entidad de sinónimos.

Las entidades complementarias se suelen utilizar tanto para aclarar aspectos de los datos de la entidad principal, como para proporcionarnos más información, además de asistir al usuario en la consulta que formula.

1.4.2.3 Las bancos de datos complementarios

Los bancos terminológicos suelen estar relacionados con otros bancos de datos que les sirven para proporcionarles información complementaria, a los cuales pueden acceder de forma autónoma con diferentes finalidades informativas (no necesariamente la información proporcionada va a ser del tipo terminológico). Los bancos de datos más representativos son los siguientes:

- Banco bibliográfico y documental: tiene registros con la información completa de las fuentes.
- Banco temático: contiene registros sobre áreas temáticas que especifican, entre otros datos, las características, el contenido de cada área y las relaciones que mantienen entre sí.

Se puede tener otros tipos de bancos complementarios pero ello depende de las necesidades que tenga cada usuario.

1.4.2.4 La estructura interna

Los bancos de datos tienen una representación externa (interfaz del usuario) y una representación interna, que está determinada por la lógica de la programación. Esta

representación interna puede tener varios niveles de percepción, todo depende del usuario que esté consultando los datos.

El sistema no va a generar los mismos resultados a las consultas que hacen entre un usuario externo al banco de datos y otro que accede con calidad de propietario de las bases de datos. En el primer caso, el usuario solo puede ver unos datos que responden a una selección previa que el sistema ha efectuado de acuerdo con los elementos que el usuario ha introducido; en el segundo caso, el usuario puede acceder a la información global del banco así como a su estructura.

1.4.3 Recuperación de la información

No se debe de perder de vista que se construye un banco terminológico para cumplir con las necesidades específicas de los usuarios y darle respuesta a sus consultas en un tiempo muy breve, por lo tanto el resultado que tienen que obtener del mismo debe de satisfacer sus exigencias. Sin embargo, la recuperación de la información está sustentada en la estructura y programación del sistema, por lo que hay que diseñar perfectamente el sistema para que cubra las necesidades para las que fue pensado, antes de comenzar con la programación.

1.4.3.1 Formas para hacer una consulta

Los usuarios pueden acceder al banco terminológico de distintas formas, todo depende de la manera en que está desarrollado el sistema. Tomando en cuenta que la relación que hay entre usuario-máquina es directa (los sistemas tienen que ser desarrollados mínimo de acuerdo con un diseño del tipo de *segunda generación*, en el que ya no es necesario un intermediario para la relación usuario-máquina), las formas en que los sistemas informáticos se suelen complementar son con menús de consulta, con manuales de ayuda automatizados los cuales explican como proceder para una consulta determinada, o bien con interfaces en lenguaje ya sea restringido o natural.

Con los menús de consulta, el usuario va haciendo su búsqueda conforme el propio menú lo va guiando. Éstas pueden variar entre cada banco. Ahora dependiendo de cuántas formas diferentes de búsqueda tenga el menú del sistema, el usuario puede obtener desde resultados con información raquítica hasta muy completa.

Los manuales automatizados, de acuerdo con las necesidades que se tengan, guían al usuario en el modo de hacer las consultas. Regularmente, esta forma de hacer consultas está basado en lenguaje restringido.

Con respecto al lenguaje restringido (el cual ha sido hasta ahora el más utilizado), está limitado a un cierto conjunto de palabras. Las palabras que se utilizan las delimita el desarrollador del sistema. Por ejemplo, si deseamos hacer una consulta para saber cuántos términos existen en idioma alemán, con el lenguaje restringido no podemos elaborar esta pregunta; probablemente lo que tengamos que hacer es marcar en un menú la opción, o escribir únicamente las palabras: términos → equivalencia → alemán, lo que provoca que

el usuario no se sienta libre en el manejo del sistema. Por otro lado, el lenguaje natural permite un diálogo más fluido entre usuario-máquina, ya que es el que usa el ser humano para comunicarse con sus semejantes y simplemente con una interfaz se establece un puente entre el usuario y el sistema.

1.4.3.2 Tipos de consulta

De acuerdo con la información que contenga el banco de datos y la potencia del sistema con que haya sido desarrollado, se establecerá la cantidad de consultas que va a aceptar el banco.

En general todo banco terminológico acepta demandas de consultas de los siguientes tipos:

- Referidas a un solo dato o a un conjunto de ellos.

  Una consulta a un solo dato puede ser del tipo:

  - Definición del término X.
  - Fuente del término X.
  - Sinónimo del término X.
  - Responsable de la inserción del término X.

  En este tipo de consultas el programa va directamente al registro y extrae la información que se le pide. Son procesos sencillos y en consecuencia muy rígidos.

  Podemos pedir más información acerca del término en cuestión, pero no se deja de hacer una consulta sencilla al banco, por ejemplo:

  - Definición, equivalencia y contexto definitorio del término X.
  - Equivalencia en inglés, alemán y francés del término X.

  Dentro de las consultas que se pueden elaborar a un conjunto de datos tenemos:

  - Lista de todos los términos que no tienen equivalencia en alemán.
  - Lista de todas las fuentes que son del año X.
  - Lista de todos los términos que proceden de la fuente X.
  - Lista de todos los responsables que ingresaron datos en la fecha X.

En este caso, el sistema ofrece información que extrajo de diferentes registros, dependiendo de la condición que le hayamos dado.

- Por una condición o una combinación de ellas.

Las consultas a un conjunto de datos, que conllevan una condición fueron ilustradas en los últimos cuatro ejemplos anteriores. Resaltando las condiciones tenemos tales como: "...que son del año X", "... que no tienen equivalencia en alemán".

Para el caso en que se tiene una mezcla de condiciones (tomando en cuenta que debe de haber 2 o más de ellas) dentro de la consulta solicitada al sistema, debemos de hacer uso de operadores boléanos u operadores lógicos para poder establecer una relación entre las condiciones. Dentro de los operadores que pueden ser utilizados tenemos *y, y no, o, mayor que, entre, and, or, not, etc.* Para ejemplificar lo anterior se tienen las siguientes consultas:

a) Lista de todos los términos elaborados por el responsable X o por el responsable Y

1ª condición:	ser del responsable X
2ª condición:	ser del responsable Y
operador de relación:	o

b) Lista de todos los términos del área temática X que no tienen equivalencia en alemán

1ª condición:	ser del área temática X
2ª condición:	que tengan equivalencia en alemán
operador de relación:	y no

1.4.3.3 Tipos de respuestas

Cuando un usuario le hace una consulta al sistema, éste genera las operaciones pertinentes para encontrar la respuesta que le parece adecuada y mostrársela al usuario. La forma de mostrarle al usuario los resultados puede ser directa o pueden ir acompañados de otras opciones para poder filtrar más la información, y así obtener datos más específicos. Por ejemplo:

Consulta: Lista de todos los términos que proceden de la fuente X.
Resultado: El banco terminológico contiene 256 registros que cumplen con ese requisito (haciendo una suposición), los cuales son mostrados al usuario. Sin embargo, pueden aparecer algunas otras alternativas al usuario:

- Mostrar los registros completos por pantalla (pueden ser uno detrás de otro o por separado).

- Imprimirlos.
- Guardarlos en otro tipo de soporte informático (ver apartado 1.4.3.4).
- Efectuar una nueva selección sobre los resultados obtenidos (en este caso se indica que criterio se va a utilizar para la selección).

1.4.3.4 Tipos de soporte informático

Se puede tener el resultado de las consultas, almacenadas en algún componente de almacenamiento o simplemente mostrarlos directamente en algún dispositivo.

- En la pantalla del monitor.
- En impresión en papel.
- En cinta magnética.
- En CD-ROM
- En disquete de 3 ½".
- En microfichas.
- En disco duro para PC.

1.4.3.5 Evaluación de la eficiencia del banco de datos

Los sistemas actuales deben incorporar un buzón electrónico, el cual permita a los usuarios manifestar sus opiniones sobre qué tan buenas fueron las repuestas que les presentó el sistema ante sus búsquedas, y además puedan hacer sugerencias con el fin de mantener actualizado e ir mejorando la calidad el banco.

Opcionalmente puede tener un sistema de evaluación automático, del rendimiento que tiene el banco, para poder estimar la eficacia del servicio que se les presta a los usuarios y caer en cuenta de qué elementos son los que necesitan perfeccionarse. Dentro de estas evaluaciones, se puede tener por ejemplo: ¿Cuántas consultas han elaborado los usuarios a las áreas temáticas del banco? ¿Cuántos registros ha ingresado al banco el responsable X?

1.5 Bancos terminológicos existentes

En este apartado, vamos a hacer referencia a los bancos terminológicos existentes en Internet que consideramos más importantes, tanto en contenido terminológico como en definiciones e idiomas.

1.5.1 EUSKALTERM

Este banco de datos terminológico es propiedad del gobierno vasco desde el 13 de julio de 2001. Ofrece información terminológica sobre el vocabulario especializado del euskera, incluye la definición de cada uno de los términos, el contexto de uso de los mismos y sus equivalencias en otros idiomas. Igualmente recoge las nuevas palabras fruto del desarrollo en los ámbitos científico y técnico. Los especialistas de cada una de las áreas, además de traductores y divulgadores, son los principales usuarios del banco terminológico del euskera que se puede consultar en Internet en la dirección http://www.euskadi.net/euskalterm, el cual es totalmente público.

1.5.2 TERMIUM

Este banco terminológico contiene en la actualidad 3.5 millones de términos, registrados en aproximadamente 1.3 millones de fichas que tratan de la traducción, la terminología y los títulos y denominaciones oficiales. Las fichas son actualizadas constantemente a fin de que reflejen la evolución del conocimiento en las esferas de la actividad humana que abarcan. Asimismo, los terminólogos mejoran regularmente el contenido de la base de datos añadiendo términos especializados en inglés, francés, español y portugués con objeto de garantizar la eficacia de las comunicaciones dentro de la función pública. Por último, cabe señalar que el contenido terminológico del banco de datos se ha visto enriquecido con la inclusión de herramientas de ayuda a la redacción en francés e inglés.

La página WEB de TERMIUM es http://www.termium.com/site/espanol/index.html. Para tener acceso al banco terminológico, es necesario darse de alta en la página de TERMIUM y ellos proporcionan el banco a prueba por 7 días.

1.5.3 EURODICAUTOM

Es un banco de datos terminológico plurilingüe del European Commission's Translation Service. Se desarrolló inicialmente para ayudar a los traductores, en la actualidad es consultado por un número cada vez mayor de responsables de la Unión Europea, además de una gran cantidad de profesionales de la lengua de todo el mundo gracias al MLIS-Project (Multilingual Information Society). La información que contiene se redacta en doce lenguas y se actualiza constantemente. Abarca una amplia visión del conocimiento humano, aunque lo fundamental son temas de interés para la Unión Europea. La base de datos contiene términos técnicos, abreviaturas, acrónimos y fraseología. El 1 de abril de 1999, la base de datos contenía más de cinco millones y medio de entradas .

El diccionario se encuentra físicamente en: http://europa.eu.int/eurodicautom/login.jsp

1.5.4 LE GRAND DICTIONNAIRE TERMINOLOGIQUE

Incluye más de 800,000 fichas de terminología y más de 3 millones de términos técnicos en francés e inglés. Cada ficha puede incluir a la vez un término en idioma francés y uno

en inglés así como sinónimos, definiciones, contextos, notas, fuentes y otras informaciones útiles.

En cuanto al banco terminológico, las búsquedas por término en francés o en inglés se ven facilitadas en el monitor gracias a una ventana que presenta el índice de los términos, empezando por el que se busca. Se pueden seleccionar hasta cien términos por vez, y se obtendrán los documentos respectivos. Cada ficha terminológica incluye el(los) ámbito(s), la parte inglesa y francesa de la ficha y sus referencias. En cada caso, se puede buscar la fuente del documento, así como precisar la búsqueda con mayor detalle, registrar el trabajo regularmente y exportar a otros ficheros o imprimir los datos. También se pueden efectuar varias búsquedas a la vez en un archivo especial, e incluir anotaciones personales.

El diccionario se encuentra físicamente en: http://www.granddictionnaire.com/_fs_global_01.htm

1.5.5 TRADOS multiterm WEB access index

Contiene más de 150,000 entradas en once lenguas y la administra la División de Apoyo Informático, Lingüístico y Documental (SILD) de la Dirección General de Traducción y Servicios Generales (DG7) de la Secretaría General del Parlamento Europeo. Su contenido incluye:

- Términos en las once lenguas oficiales de la Unión Europea.
- Términos en latín (nombres científicos).
- Términos en lenguas exteriores a la Unión.
- Acrónimos y abreviaturas de gran cantidad de términos.

Los términos de la base abarcan los siguientes campos temáticos: terminología europea tal y como aparece en el boletín oficial de la Unión Europea, incluyendo la terminología relativa a los distintos proyectos y programas, reglamento del parlamento europeo, organigrama del parlamento europeo, gobiernos locales y regionales, partidos políticos de los estados miembros, informática y telecomunicaciones: medicina, salud, SIDA: botánica y zoología; política social y derecho de asilo; educación, formación, diplomas y calificaciones profesionales; ecología y contaminación; transporte y control aéreo.

El diccionario se encuentra físicamente en: http://muwa.trados.com

1.5.6 Terminology forum

Terminology forum contiene un directorio de diccionarios que lleva por nombre *Terminology collection*. Está dividido en dos partes:

- WORD ON LINE (Diccionarios de lengua general): Aquí se puede encontrar una lista de diccionarios que se encuentran en Internet y en Gopher los cuales son gratuitos. Dentro de la lista de diccionarios puedes encontrar de tipo monolingües, bilingües y plurilingües.

- TERM ON LINE (Glosarios de lengua especializada): En esta parte se puede encontrar una lista de glosarios de diferentes campos, como por ejemplo computación. Están clasificados en primer lugar de acuerdo al campo y en segundo lugar al (los) tipo(s) de idioma(s).

La dirección de la página WEB de Terminology forum es: http://www.uwasa.fi/comm/termino/

El directorio se encuentra en la dirección de Internet: http://www.uwasa.fi/comm/termino/collect/index.html

1.5.7 Base de terminologie

Es un banco terminológico plurilingüe el cual fue elaborado por el Consejo Internacional de la Lengua Francesa. Los términos se encuentran en francés, inglés, español y alemán. Contiene definiciones de los términos únicamente en idioma francés. El diccionario se puede consultar en Internet en la dirección: http://www.cilf.org/bt.fr.html

1.5.8 TIS (Terminological Information System)

TIS son las siglas para el Sistema de Información Terminológico de la Secretaría General del Consejo de la Unión Europea. Este sistema fue diseñado como una herramienta de trabajo para los traductores del consejo, y ha permitido proporcionar una base de datos de términos y de abreviaturas en los idiomas de las comunidades al servicio de la terminología.

El banco se encuentra disponible en los idiomas danés, holandés, inglés, finlandés, francés, alemán, griego, irlandés, italiano, portugués, español, sueco y latín. Los temas incluyen agricultura, el ambiente, la energía atómica, la armonización de la legislación y la transformación de los alimentos.

El sistema contiene actualmente alrededor de 200,000 entradas, las cuales 45% de ellas contienen entradas en tres o más idiomas. El crecimiento de la base de datos actualmente es de 4,000 traducciones por mes.

El banco se encuentra en: http://tis.consilium.eu.int/utfwebtis/frames/introfsEN.htm

1.5.9 ILOTERM

Es una base de datos terminológica en cuatro idiomas que contiene 430,000 registros aproximadamente. Está a cargo de la Unidad de Terminología y Referencias, que pertenece al Servicio de Documentos Oficiales (OFFDOC).

Su finalidad principal es brindar soluciones a los problemas terminológicos que se plantean en los ámbitos social y laboral. Los términos se incorporan en inglés, con sus equivalentes en francés, español y, en su caso, alemán. La base de datos incluye asimismo registros (como máximo en cuatro idiomas) relativos a la estructura y a los programas de la OIT (Organización Internacional del Trabajo), los nombres oficiales de instituciones internacionales, organismos nacionales, organizaciones de empleadores y de trabajadores, así como a los títulos de reuniones e instrumentos internacionales.

Los términos y sus equivalencias proceden de numerosas fuentes: instrumentos, informes y estudios de la OIT, publicaciones periódicas, de diarios y de otras instituciones internacionales, diccionarios y glosarios técnicos, así como aportaciones de expertos y de lingüistas.

La recopilación de datos se ha llevado a cabo a lo largo de varios años, pero no se considera en modo alguno exhaustiva. ILOTERM se actualiza periódicamente, y es posible buscar los términos y las abreviaciones en cualquiera de los cuatro idiomas.

La ubicación del banco es: http://www.ilo.org/public/spanish/support/lib/dblist.htm

1.5.10 Telecommunication Terminology Database (TERMITE)

Los usuarios tienen la posibilidad de tener acceso al banco de datos terminológico de ITU (International Telecomunication Union). TERMITE contiene todos los términos que aparecieron en glosarios impresos en la ITU desde el año 1980 hasta nuestros días, así como las entradas más recientes referente a las diversas actividades de la unión, en total son unas 59,000 entradas aproximadamente. El enlace para acceder a este banco terminológico es: http://www.itu.int/search/wais/Termite/

1.5.11 IMF TERMINOLOGY

Este banco terminológico contiene 4,500 términos aproximadamente, los cuales son útiles a los traductores que trabajan con el material del IMF (International Monetary Fund). Proporciona los términos y su equivalencia en alemán, inglés, francés, español y portugués.

El banco de datos incluye: palabras, frases, y los títulos institucionales encontrados comúnmente en documentos del IMF, en áreas tales como economía y actividades bancarias, finanzas públicas, balance de pagos, y desarrollo económico. El banco se encuentra en: http://www.imf.org/external/np/term/index.asp

Los bancos terminológicos antes mencionados fueron considerados los más representativos de la gran variedad existente en Internet. El motivo de considerar únicamente a los de este tipo, radica en su fácil acceso y su disponibilidad gratuita en su mayoría de ellos.

Se observó que existe una gran variedad de bancos que pueden contener equivalencias, acrónimos, abreviaturas, definiciones, pueden ser monolingües o plurilingües, etc., pero no implica que todos contengan lo mismo, dado que son diversos en sus áreas de especialidad, idiomas de consulta, tipos de búsquedas, etc.

Como se ha indicado a lo largo del presente capítulo, los bancos terminológicos se encuentran diseñados para cumplir con las necesidades que tiene cada usuario, y no es posible utilizar alguno de los mencionados anteriormente porque ninguno considera algunos campos que son vitales para cumplir con los requerimientos que tiene el Grupo de Ingeniería Lingüística. El motivo de lo antes citado se explica a detalle en el capítulo 3.

Capítulo 2. FAQ de requerimientos para el banco terminológico del GIL

Al desarrollar un sistema se deben de tener bien cimentadas las ideas hacia donde se pretende llegar, esto es, ¿para qué nos va a servir? y ¿cómo lo vamos a hacer? En este sentido, en el presente capítulo identificaremos los requerimientos necesarios para desarrollar el banco terminológico.

De entrada es necesario hacer un análisis exhaustivo de las necesidades que se tienen, ya que éstas son las bases para la elaboración de un buen sistema y que cumpla con los objetivos propuestos. Para aclarar todas las dudas que se tengan al respecto es benéfico hacernos ciertas interrogaciones. Para tal caso Ventura Miranda[1] nos presenta una serie de preguntas frecuentes (FAQ, Frequently Asked Questions), útiles para la obtención de requerimientos que contestaremos con el fin de cimentar las ideas que se tienen.

Las preguntas están clasificadas y ordenadas, de tal manera que se lleve un orden en los requerimientos y no se deje algo en el olvido. La clasificación es en 7 series de preguntas: contexto, problemática, volumen y tipo de información, perfil y participación de los usuarios, características deseadas, infraestructura y tipología.

2.1 Contexto

Estas preguntas abordan cuestiones como: ¿quién y cómo surge la iniciativa del proyecto?, ¿de dónde provendrá el apoyo?, ¿el desarrollo del sistema será factible?, así como ¿quién beneficiará la elaboración del sistema? En otras palabras, las preguntas enfocadas al contexto nos darán un panorama más amplio de todo el departamento estratégico que rodea la elaboración del banco terminológico.

¿Quiénes están detrás de la iniciativa de trabajo?

Nuestra investigación se encuentra apoyada por el Grupo de Ingeniería Lingüística (GIL), en el seno del Instituto de Ingeniería de la UNAM[2].

El GIL representa un grupo de investigación en la que dos áreas, al parecer alejadas, se plantean un sentido de unidad e interdependencia para formar un solo núcleo. Estas áreas son la lingüística y la ingeniería. La ingeniería lingüística es un área interdisciplinaria de investigación aplicada al desarrollo de sistemas computacionales para reconocer, interpretar y generar lenguaje humano. Existe una correspondencia biunívoca, de forma que la lingüística permite la creación de modelos en lenguaje natural que puedan ser utilizados por los sistemas computacionales, mientras que la ingeniería permite el desarrollo de sistemas que puedan resolver las necesidades específicas planteadas por los problemas lingüísticos.

[1] Ventura (2002)
[2] La información del GIL se puede consultar en http://iling.torreingenieria.unam.mx

El grupo se conformó, primero, con el fin de crear una base de conocimiento relativa y concerniente a esta área de trabajo, y segundo, de formar personal especializado y comprometido con el estudio y desarrollo de las diversas áreas que ésta ofrece. Su interés radica en la realización de proyectos que superen las necesidades y los problemas presentados para el procesamiento del lenguaje natural, incluyendo el desarrollo de aplicaciones específicas que sirvan a las diferentes áreas con las que el GIL interactúa.

¿De dónde surgió la iniciativa, qué áreas intervienen y con qué apoyo se cuenta?

La elaboración del banco terminológico nace a partir de uno de sus proyectos principales que desarrolla el GIL, alrededor del cual giran diferentes líneas de investigación. Este proyecto esta encaminado a elaborar un diccionario que realice búsquedas onomasiológicas[3], esto es, un diccionario que permita la búsqueda de términos a partir de la descripción del concepto mediante el uso de lenguaje natural. El diseño y creación del diccionario onomasiológico comprende fases bien definidas: adquisición de datos (extracción terminológica y extracción conceptual), creación de bases de datos y captura de información, determinación de paradigmas semánticos, diseño del motor de búsqueda y diseño de la interfaz del usuario (figura 2-1)[4]. El trabajo en conjunto de este diccionario se ve beneficiado (como se verá en la siguiente pregunta) del banco terminológico.

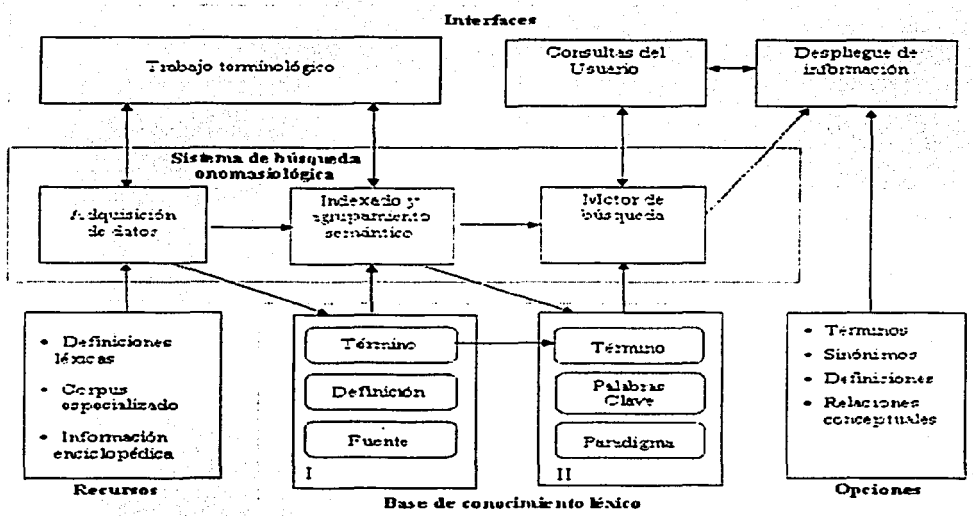


Figura 2-1. Arquitectura del diccionario onomasiológico

[3] Sierra (1999)
[4] Sierra, Castillo, Reyes y Alarcón (2001)

Gracias al apoyo del Instituto de Ingeniería y con el patrocinio del Consejo Nacional de Ciencia y Tecnología y de la propia UNAM, el GIL ha podido sustentar el desarrollo de la presente investigación.

¿Qué beneficios se esperan de la solución?

La elaboración del presente banco terminológico ayudará a mantener almacenada una gran cantidad de información en un solo lugar, así como tener acceso a ella de un modo fácil y eficiente. El conjunto de toda la información forma las bases de conocimiento del banco y con ellas se beneficiarán diferentes líneas de investigación que se están siguiendo dentro del GIL:

- En la adquisición de datos, para alimentar las bases de conocimiento del banco terminológico. El punto aquí es buscar y extraer la terminología a partir de textos especializados y de diccionarios existentes. Una vez obtenida la terminología se sujeta a revisión por expertos en la materia, los cuales dan el visto bueno de que se acepta la terminología obtenida para introducirla en el banco. Las áreas temáticas con las que cuenta el banco son: física, lingüística, desastres, ingeniería lingüística, ingeniería, metrología, fenómenos destructivos y sexualidad. Para la extracción terminológica de manera automática a partir de textos especializados[5], se usaron técnicas de extracción terminológica, por lo que se utilizó un programa comercial denominado WordSmith[6], que permite comparar las frecuencias de las palabras en dos textos de diferentes temáticas (el texto en cuestión y otro de un área completamente diferente).

- Se ha observado que cuando un autor introduce un término nuevo que no es conocido del todo por sus lectores, normalmente utiliza una serie de patrones sintácticos y tipográficos en donde se incluye la definición del término. En caso de existir un inventario de estos patrones, entonces sería posible desarrollar una herramienta capaz de extraer automáticamente los contextos definitorios, facilitando el trabajo terminológico y permitiendo una clasificación más rápida de los posibles términos y sus definiciones. Para facilitar la identificación de conceptos en un área determinada a partir de textos especializados, el GIL desarrolla, en colaboración con la Universidad Pompeu Fabra[7] un método de identificación de patrones recurrentes para la extracción automática de contextos definitorios[8]. Al término con su(s) respectiva(s) definición(es) englobados en conjunto en un texto de especialidad, se le denomina contexto definitorio.

- Para que un proceso de búsqueda onomasiológica entregue al usuario un resultado útil es preciso contar con una base de redes de conocimiento. Para este fin, el GIL trabaja en la creación de grupos de paradigmas semánticos, los cuales se emplean en un proceso de búsqueda onomasiológica a partir del lenguaje natural. Los

[5] Reyes (2002)
[6] http://www.oup.com/elt/global/isbn/6890/
[7] http://www.upf.es/
[8] Alarcón y Sierra (2002)

paradigmas semánticos son grupos de palabras clave que pueden ser utilizadas como sinónimos en ciertos contextos. Para la determinación de estos paradigmas se sigue una metodología de agrupamiento semántico[9] a partir del alineamiento de definiciones.

En las investigaciones en curso será benéfico contar con el banco terminológico, ya que en principio todas las líneas de investigación están trabajando sobre términos y definiciones, además de algunos otros campos como los contextos definitorios. De igual modo, es preciso contener almacenado en algún lugar todas las fuentes de donde ha provenido la información que contendrá el banco.

¿Cómo se evaluaría el éxito de la solución?

En primer lugar, el hecho de mantener toda la información terminológica con sus respectivas definiciones de las diferentes áreas temáticas ya es un indicio de que el sistema es funcional. En segundo lugar, los campos adicionales que se introducirán en el banco terminológico, con el fin de apoyar las investigaciones en proceso, darán un valor agregado al banco. Por último, el hecho de que el sistema se elaborará por medio de protocolo TCP/IP, proporcionará a los usuarios un gran beneficio, ya que podrán utilizar la herramienta desde cualquier lugar que cuente con una conexión a Internet.

2.2 Problemática

En esta sección nos abocaremos a analizar una serie de preguntas que nos permitan hallar todos los problemas y deficiencias que se tienen en el GIL, en el cual involucren la necesidad de un banco terminológico.

¿Qué problemas existen actualmente?

La información que ha recopilado el GIL se ha mantenido por separado, es decir, se ha estado guardando de tal manera que ha sido catalogada como diccionarios especializados aunque aislados. En consecuencia, los integrantes del grupo tienen la necesidad de poseer un sistema que contenga toda la información recopilada en un solo lugar, para que puedan obtener una respuesta a sus peticiones de información de una forma agradable, rápida, segura, desde cualquier lugar en el que se encuentren. El hecho de contar con la información de esta manera, los usuarios podrán utilizar el banco como una herramienta para las líneas de investigación que se están siguiendo dentro del grupo.

¿Por qué existe el problema?

Debido a una serie de requerimientos que se han suscitado, por el hecho de que se han abierto más líneas de investigación en el grupo, el GIL se ve en la necesidad de tener un sistema donde pueda almacenar la información compilada, de una forma integral y que sirva de herramienta para las diferentes líneas de investigación.

[9] Castillo (2002)

¿Cómo se resuelve actualmente?

La solución a los problemas encontrados es la elaboración de un banco terminológico en donde se encuentre toda la información que el GIL ha recopilado, y que con las búsquedas que presente el sistema genere una respuesta que resuelva las carencias de información de los usuarios.

El GIL contará con un sistema electrónico que sirva para la captura de los términos y definiciones para las bases de conocimiento. Con el sistema de base de datos desarrollado, una vez que sea aprobado por los expertos, se podrá capturar la información correspondiente a la terminología con sus respectivas definiciones, las cuales se extraerán a partir de textos especializados y de diccionarios existentes. El sistema será implantado en una plataforma en Internet, a fin de que pueda usarse por diferentes capturistas en distintas máquinas. El sistema será capaz de proporcionar un acceso fácil, confiable y seguro a la base de datos a través del Web, permitiendo que sólo los usuarios autorizados alimenten la base de datos de manera simultánea, sin restricción de acceso, a la vez que permitirá ver la información, únicamente para consulta, a toda persona que visite la página Web.

2.3 Volumen y tipo de información

Después de haber encontrado una solución a los problemas y las carencias que el GIL tiene hasta ahora, podemos pasar a definir ¿qué tipo de información contendrá el banco? ¿quién la va a proveer? ¿cómo se encuentra almacenada, distribuida, etc.? ¿cuánto espacio va a ocupar la información?

¿Qué tipo de información se va a procesar?

Toda la información que el GIL ha recopilado, proviene de textos especializados y diccionarios, por esta razón el banco terminológico contendrá información específica. Esta información incluirá aspectos como: términos, definiciones, contextos definitorios, sinónimos, equivalencias, fuentes (año, lugar de publicación, editor, editorial), áreas temáticas, responsables y fechas tanto de inserción como de modificación de la información.

¿Quién provee dicha información?

Como se ha mencionado anteriormente, la información parte de textos especializados y diccionarios para después someterlos a la extracción terminológica, así como la aprobación por parte de los expertos y finalmente almacenarla en el sistema con ayuda de capturistas de datos.

¿Cómo se encuentra almacenada, representada o distribuida la información?

La información se encontrará residente en bases de datos relacionales, agrupada en 8 diferentes campos: física, lingüística, desastres, ingeniería lingüística, ingeniería,

metrología, fenómenos destructivos y sexualidad. La información estará distribuida en 9 tablas diferentes, dicha distribución será explicada a detalle en el capítulo 3.

¿Cuántos mega bytes ocupará la información?

Hasta el momento, la información que se ha almacenado para el banco terminológico tiene un total de 32.17 mega bytes distribuidos de la siguiente manera (Tabla 2-1):

Tabla 2-1. Distribución de la información del banco terminológico

Base	Mega bytes de almacenamiento
Física	3.32
Lingüística	9.30
Desastres	7.21
Ingeniería Lingüística	2.35
Ingeniería	1.97
Metrología	2.46
Fenómenos destructivos	2.63
Sexualidad	2.93

¿Cuántos archivos se necesitarán?

Dado que se están utilizando bases de datos, no habrá necesidad de generar archivos adicionales en el desarrollo del banco terminológico, sino solamente los que genere el propio sistema manejador de bases de datos.

2.4 Perfil y participación de los usuarios

Esta sección está dedicada a conocer quién va a utilizar el sistema desarrollado, e igualmente considerar si son necesarias algunas aptitudes especiales para el manejo del banco. o si se requiere capacitar a los usuarios.

¿Quiénes utilizarán la solución?

Dado que es un sistema que será desarrollado para las necesidades específicas del Grupo de Ingeniería Lingüística, los usuarios potenciales del banco serán todos los integrantes del grupo, los que colaboran con éste (la Universidad Pompeu Fabra y el Colegio de México[10]) así como los expertos calificados. Por otro lado, podrán tener acceso al banco las personas que estén interesadas en consultar información de carácter científico y/o tecnológico, o que tengan la necesidad de poseer una base terminológica en cualquier área temática, dependiendo de las propias necesidades del solicitante. Dentro de los usuarios que no tienen ninguna relación con el GIL y que les debe resultar atractivo hacer uso del banco encontramos a los especialistas de la comunicación, la información y la documentación; los expertos en lexicografía, terminología y traducción; ingenieros lingüistas y lingüistas computacionales; editores,

[10] http://www.colmex.mx

profesores de lengua, investigadores en lingüística aplicada; profesores de áreas especializadas; organizaciones de la sociedad civil con trabajos relacionados a los temas; escuelas y estudiantes de secundaria, preparatoria y universidad.

¿A qué área pertenecen?

Las áreas que están involucradas en el desarrollo del banco terminológico son multidisciplinarias, por lo que puede parecer que no hay relación entre todas ellas. Sin embargo, para el campo de investigación que la solicitó sí lo es y todas las áreas en su conjunto son de suma importancia para el GIL. Las áreas involucradas son física, lingüística, desastres, ingeniería lingüística, ingeniería, metrología, fenómenos destructivos, sexualidad y está en proceso de investigación el área de arabismos.

¿Cuántos usuarios utilizarán la solución?

Inicialmente se tienen registrados a 27 usuarios que harán uso de la solución en su totalidad, aparte hay otros usuarios adicionales que trabajan en alguna base en particular; sin embargo, no se tiene un límite en la cantidad de usuarios que van a acceder al banco, ya que por el hecho de estar implantado en Internet podrán acceder una infinidad de ellos con previa autorización o en su caso únicamente para consulta del mismo.

¿Cuál es el perfil del usuario?

No se tiene un perfil definido, ya que los integrantes del GIL y la gente que colabora con éste son de diferentes perfiles. Sin embargo, todos ellos coinciden en un mismo fin: utilizar el banco terminológico como una herramienta que les va a ayudar a resolver problemas específicos.

¿Cuál es la formación profesional del usuario?

La formación de los usuarios puede ser variable, dado que el banco terminológico contiene distintas áreas temáticas, puede ser que alguna de ellas sea de interés para usuarios con un nivel de estudios de secundaria o para usuarios con un nivel de doctorado.

¿Cuál es la formación de cómputo del usuario?

Deben de tener los conocimientos básicos de cómputo y navegación en Internet, para que pueda consultar el banco terminológico por la Web.

¿Tienen experiencia en este tipo de aplicaciones?

No es necesario que cuenten con algún tipo de experiencia, pero el hecho de que ya hayan usado algún diccionario electrónico, puede ayudarles con el manejo del banco terminológico.

¿Qué tipo de capacitación será necesaria?

No es indispensable contar con capacitación alguna, ya que el propio banco los llevará de la mano para que los usuarios puedan generar sus consultas. Sin embargo, el banco terminológico tendrá su propia ayuda.

2.5 Características deseadas

En este apartado se definirán las características principales que se requieren para el sistema, los requerimientos de seguridad que serán necesarios implantar, así como el soporte del producto.

¿Cuáles son las características principales que se esperan del sistema?

El sistema deberá mostrarnos todos los términos con los que cuenta en su base de conocimiento, así como la información que le corresponde a cada uno de ellos, en interfaces que sean amigables para el usuario. Las búsquedas que almacenará el sistema tienen que satisfacer las necesidades de los usuarios, por ello se tendrá que implantar en una plataforma de Internet para tener acceso al banco desde cualquier lugar.

¿Qué funcionalidad debe de proporcionar?

Para que el banco terminológico sea funcional, deberá presentar los términos con su respectiva información (definición, contexto definitorio, etc.) en interfaces amigables. Lo más importante que tendrá que hacer el banco es cumplir con las necesidades que se tienen en las diferentes líneas de investigación que está llevando a cabo el GIL, mencionadas en las preguntas del contexto (2.1).

¿Quién dará soporte al producto?

El sistema estará respaldado por un experto en administración de servidores Web y bases de datos. La persona destinada a desempeñar este cargo tendrá conocimientos de bancos terminológicos para proporcionar un servicio eficiente. Por otro lado el sistema contará con soporte en línea directamente del administrador del sistema, para cualquier comentario o sugerencia por parte del usuario.

¿Cuáles son sus requerimientos de seguridad?

El banco tendrá un sistema de autentificación de usuarios, con el fin de contar con un registro de quién es la persona que ha entrado y salido. Por otro lado, dependiendo el usuario que acceda al sistema serán el tipo de privilegios que va a tener dentro del mismo; esto es porque un usuario que necesite información de tipo administrativa sobre el sistema, no tendrá los mismos permisos que un usuarios que está trabajando en introducir información al banco y éste no tendrá los mismos permisos que el usuario que necesita únicamente información con el fin de documentarse sobre algún tema en específico.

Con el fin de mantener protegida la información, se generarán respaldos diarios de todas las bases de datos existentes en el banco terminológico. El motivo por el cual se realizará esto, es porque día con día se estará actualizando el banco a través de la gente que trabaje en la adquisición y llenado del mismo.

¿Cómo será distribuido el software?

El banco terminológico estará disponible en Internet, implantado en una página Web, de forma que pueda ser consultado desde cualquier lugar del mundo. El acceso público al banco será únicamente para consultar la información. En dado caso que el usuario tenga la necesidad de introducir o hacer alguna modificación en la información podrá realizarla bajo previa autorización.

2.6 Infraestructura

Para implantar un sistema se necesitan recursos tanto económicos como de infraestructura. En esta sección nos dedicaremos a hablar de la infraestructura con la que cuenta el GIL para desarrollar el banco terminológico.

¿Qué plataforma se usará?

La plataforma a utilizar será Windows 2000 Server. La razón básica por la cual se utilizará esta plataforma radica en que el manejador de las bases de datos será Microsoft SQL Server. Se usará Microsoft SQL Server[11] debido a que proporciona servicios seguros y escalables de almacenamiento de datos relacionales, incluye completas capacidades de administración y permite obtener una gran disponibilidad de recursos. Por otro lado, permite un fácil acceso a los datos a través del Web. Del mismo modo, se puede utilizar HTTP para enviar consultas a la base de datos, realizar búsquedas de texto en documentos almacenados en la base de datos y ejecutar consultas a través del Web con lenguaje natural.

¿Existen planes para plataformas futuras?

En un principio se desarrollará el sistema una plataforma Windows, para posteriormente implantar el banco terminológico en plataforma Linux debido por una parte a que los sistemas que se estarán desarrollando dentro del GIL serán orientados a utilizar esta plataforma, y por otra parte Linux es un sistema operativo que es gratuito, cuenta con un buen sistema de seguridad para la información y es portable a cualquier tipo de computadora.

¿Con qué recurso de hardware y software cuentan actualmente?

Por el lado del software actualmente se cuenta con un servidor que tiene instalado el sistema operativo Windows 2000 Server, así como un manejador de bases de datos

[11] http://www.microsoft.com/latam/sql/evaluation/overview/default.asp

SQL Server 2000. Desde el punto de vista del hardware, este servidor cuenta con un procesador Pentium III con 256 MB en memoria RAM, disco duro de 40 GB.

¿Existen otros sistemas o aplicaciones que sean relevantes para este proyecto?

El banco terminológico está diseñado para ser independiente, lo cual implica que no tendrá la necesidad estar ligado a algún otro sistema para que funcione; sin embargo, el propio banco es parte integral del sistema de búsqueda onomasiológica que está desarrollando el GIL.

2.7 Tipología

En el capítulo 1 se escribió todo acerca de los bancos terminológicos; ahora es el momento de analizar el banco terminológico que se elaborará dentro del GIL. Basándonos en tal capítulo, contestaremos las siguientes preguntas que nos darán un panorama más amplio del tipo de banco y la clasificación en la cual estará inmerso el banco.

¿A qué generación pertenece el banco terminológico?

Es un banco de segunda generación, debido a los siguientes aspectos:

- Los términos serán seleccionados en función de las necesidades que tengan los usuarios, esto es, la aprobación de la terminología dependerá de los expertos en el área en cuestión y esto nos va a garantizar que la información que esté contenida en el banco sea veraz, confiable y de actualidad.

- La arquitectura del banco estará diseñada de forma que se tenga una lógica y una estructura flexible y efectiva para cumplir con las necesidades de los usuarios.

- El tiempo que se requerirá para una consulta y para la recuperación de información será mínimo.

- Los usuarios realizarán sus consultas de manera autónoma y accederán al banco terminológico vía Internet.

¿Cuál es la clasificación del banco?

Para contestar esta pregunta están involucrados 7 puntos a considerar, que en su conjunto conforman la clasificación del banco terminológico. La clasificación se delimita por:

- Sus objetivos: El banco será descriptivo ya que la información contenida en éste será tal y como la presenta el autor, es decir, no se lleva alguna norma o regla para delimitar el término o en su caso definirlo.

- La organización de sus datos: El banco será organizado por el término, para lo cual las consultas partirán del término y de esta manera nos guiará hasta su definición. Comúnmente a este tipo de organización de datos se le conoce como búsqueda semasiológica.

- Su temática: El banco será de temática especializada general porque contendrá información sobre distintas áreas, con una gran cantidad de términos en campos diversos.

- El interés prioritario de los datos que contienen: El banco será de términos contenidos en documentos, es decir, los términos, sus respectivas definiciones e información adicional se encuentran inmersos en los textos y diccionarios especializados.

- El número de idiomas de las informaciones terminológicas: El banco será monolingüe con equivalencias en otros idiomas. La información completa del término se presentará sólo en un idioma y se asignarán las equivalencias denominativas en otros idiomas.

- El hardware del banco: El banco será desarrollado en una PC, especializado en cualquier área temática, conteniendo información más actual e innovadora. El área dependerá de las necesidades que se le vayan presentando al GIL.

- El tipo de banco: El banco será de segunda generación el cual ya se explicó en la pregunta anterior.

Capítulo 3. Bases de Datos

Los sistemas de bases de datos son diseñados para manejar grandes cantidades de información; la manipulación de los datos involucra tanto la definición de estructuras para el almacenamiento de la información como la provisión de mecanismos para el uso de la misma. Además, un sistema de base de datos debe tener implantados mecanismos de seguridad que garanticen la integridad de la información, a pesar de caídas del sistema o intentos de accesos no autorizados.

3.1 Definiciones y conceptos básicos

Para tener una mejor comprensión de los conceptos que se presentan en este apartado, es necesario conocer una serie de conceptos básicos, los cuales se explican enseguida. El presente apartado está basado en Olguín (1977), Silberschatz, Korth y Sudarshan (1998) y adicionalmente se indicará la fuente en el caso que no sea así.

Dato: Es un conjunto de caracteres con algún significado, los cuales pueden ser numéricos, alfabéticos, o alfanuméricos.

Información: Se trata de un conjunto de datos ordenado de tal manera que puedan ser manejados según las necesidades del usuario. Para que un conjunto de datos pueda ser procesado eficientemente y presentar su información, primero se deben guardar en archivos con una estructura lógica, es decir, que el almacenamiento de los datos tenga coherencia para que la representación de los datos genere información precisa.

Software: Es un conjunto de programas, documentos, procedimientos y rutinas asociados con la operación de un sistema de cómputo. En otras palabras, consiste simplemente en el conjunto de instrucciones individuales que se le proporciona al microprocesador para que pueda procesar los datos y generar los resultados esperados.

Hardware: Son todos aquellos componentes físicos y tangibles de una computadora. El hardware realiza 4 actividades que son fundamentales en las computadoras: entrada, procesamiento, salida y almacenamiento secundario de la información.

Base de datos: Es una colección o depósito integrado de datos almacenados en soporte secundario (no volátil), junto con un paquete de software para gestión de dicho conjunto de datos, de tal modo que se pueda controlar el almacenamiento de datos redundantes. Los datos son independientes de los programas que los usan, lo que permite que sean compartidos por diferentes usuarios y aplicaciones.[1] Se almacenan los datos y todas las relaciones existentes entre ellos, con el fin de que se pueda acceder a éstos por diferentes caminos.

Usuarios de las bases de datos: Podemos definir a los usuarios como toda persona que tenga cualquier tipo de contacto con el sistema de base de datos desde que éste se

[1] Noyano y Fernández (1998)

construye hasta su uso final. Los usuarios que acceden a una base de datos pueden ser de tres tipos:

- Administrador de la base de datos (DBA, Data Base Administrator): Es el usuario más importante debido a que se encarga de diseñar y modificar la estructura de la base de datos, administrar permisos, derechos y creaciones sobre nuevas bases de datos, así como de asignar el espacio en disco y las prioridades del procesamiento de los datos.

- Programadores de las aplicaciones: Son los expertos en computación que interactúan con el sistema por medio de llamadas en DML (Lenguaje de Manipulación de Datos), las cuales están incorporadas en un programa elaborado a partir de un lenguaje de programación (Perl, PHP, C, C++, etc.)

- Usuarios finales: Estos interactúan con el sistema invocando a uno de los programas de aplicación permanentes que se han elaborado anteriormente en el sistema de base de datos, por medio de una interfaz gráfica. Este tipo de usuario utiliza la base de datos sin saber nada del diseño interno de la misma.

Sistema Manejador de Base de Datos (DBMS, Data Base Management System): La parte más importante de un sistema de bases de datos es el software que se utiliza para la gestión de la información, el cual se le denomina *Sistema Manejador de Bases de Datos*. Un DBMS es una colección de numerosas rutinas de software interrelacionadas, cada una de las cuales es responsable de una tarea específica. Todas las peticiones de acceso a la base se manejan por medio del DBMS, por lo que este software funciona como intermediario entre los usuarios y la base de datos.

Funciones del DBMS: La función del DBMS no se limita solo a permitir mediante la definición y manipulación de datos el diálogo entre los usuarios y la base de datos. Brinda, además, mecanismos que permiten controlar la concurrencia de usuarios, así como la seguridad e integridad de la base datos y su mantenimiento en buen estado, incluso después de que haya ocurrido una falla en el sistema, sea ésta provocada por software o por hardware. Las funciones primordiales de un DBMS son:

- Crear y organizar las bases de datos.

- Establecer y mantener las trayectorias de acceso a la base de datos de forma que los datos puedan ser accedidos rápidamente.

- Manejar los datos de acuerdo con las peticiones de los usuarios.

- Registrar el uso de las bases de datos.

- Respaldar y recuperar la información, lo que consiste en contar con mecanismos implantados que permitan la recuperación fácilmente de los datos en caso de ocurrir fallas en el sistema de base de datos.

- Controlar la interacción entre los usuarios concurrentes para no afectar la inconsistencia de los datos.

- Contar con mecanismos de seguridad e integridad, que permitan el control de la consistencia de los datos evitando que se vean perjudicados por cambios no autorizados o previstos.

Lenguaje de definición de datos (DDL, Data Definition Language): Es una serie de términos o definiciones que se expresan en un lenguaje especial con las cuales se delimitan o declaran los objetos de la base de datos.

Lenguaje de manipulación de datos (DML, Data Manipulation Language): Se trata de un conjunto de expresiones que permiten manipular los datos dentro de la base de datos (insertar, recuperar, eliminar o modificar), además de regular el acceso de los usuarios a los datos. Existen básicamente 2 tipos de lenguajes de manipulación de datos:

- Lenguaje con procedimientos (álgebra relacional): Nos indica de qué manera se debe obtener información.

- Lenguaje sin procedimientos (cálculo relacional): Nos señala qué tipo de información obtenemos.

Lenguaje de control de datos (DCL, Data Control Language): Contiene elementos útiles para trabajar en un entorno multiusuario, en el que es importante la protección de los datos, la seguridad de las tablas y el establecimiento de restricciones en el acceso, así como elementos que coordinen el proceso de compartir los datos por parte de usuarios concurrentes, previniendo que no interfieran unos con otros.

Modelo de datos: El modelo de datos es una colección de herramientas conceptuales para describir los datos, las relaciones que existen entre ellos, la semántica asociada a los datos y las restricciones de consistencia. El modelo de datos es una combinación de tres componentes:

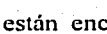
- Una colección de estructuras de datos, que son los bloques constructores de cualquier base de datos que conforman el modelo.

- Una colección de operadores o reglas de inferencia, los cuales pueden ser aplicados a cualquier instancia de los tipos de datos listados en una estructura de datos, así como para consultar o derivar datos de cualquier parte de estas estructuras en cualquier combinación deseada.

- Una colección de reglas generales de integridad, que definen un conjunto de estados consistentes (en ocasiones, son expresadas como reglas de insertar-actualizar-borrar).

Modelo relacional: Existen otros tipos de modelado (p.e., jerárquico, de red y entidad-relación), con los que se puede hacer el análisis de las bases de datos. En nuestro caso, utilizaremos el modelo relacional ya que es más fácil de comprender y analizar, además de que es el más usado por los desarrolladores de bases de datos.

Un sistema de información de bases de datos relacional se organiza en forma de tablas, las tablas se organizan en renglones y columnas, cada renglón se denomina registro y el registro contiene información referente a una instancia (estado que presenta una base de datos en un tiempo dado). Cada columna se denomina campo y la información es de un solo tipo para todas las instancias.

Modelo Entidad-Relación: El modelado de datos entidad-relación se basa en una percepción del mundo real la cual está formada por objetos básicos, llamados entidades, y por relaciones entre estos objetos (uno a muchos  , muchos a uno  y muchos a muchos ), además de las características de estos objetos llamados atributos.

Normalización de bases de datos: Las reglas de normalización están encaminadas a eliminar redundancias e inconsistencias de dependencia en el diseño de las tablas en una base de datos. Existen 4 pasos progresivos para normalizar una base de datos:

- Primera forma normal (1FN): Una relación está en 1FN sí y solo si todos los dominios simples subyacentes contienen solo valores atómicos, es decir, tienen un solo valor. Para que una relación lo sea realmente, debe contar con todos sus valores atómicos.

- Segunda forma normal (2FN): Una relación está en 2FN sí y solo si está en 1FN y todos los atributos no clave dependen por completo de la llave primaria. La *llave primaria* es una llave candidato que se usará para identificar los renglones de la tabla. Una *llave candidato* es una superllave mínima y pueden existir varias al mismo tiempo en una tabla, y a su vez una *superllave* es un conjunto de uno o más atributos que, considerados conjuntamente, nos permiten identificar de manera única a cada entidad dentro del conjunto de entidades.

- Tercera forma normal (3FN): Una relación esta en 3FN sí y solo si está en 2FN y no existen dependencias transitivas entre los atributos, esto es, nos referimos a dependencias transitivas cuando existe más de una forma de llegar a un atributo de una relación.

- Forma normal Boyce Cood (FNBC): Una relación R está en FNBC sí y solo si está en 3FN, además de que cada determinante es una llave candidato. Siendo x e y atributos de R, a x se le denomina determinante, ya que x determina el valor de y.

3.2 Obstáculos de los sistemas de bases de datos

Para realizar un buen banco terminológico, es necesario superar una serie de obstáculos que normalmente se presentan en los sistemas de base de datos[2]:

3.2.1 Redundancia e inconsistencia de datos

Puesto que los archivos que mantienen almacenada la información son creados por diferentes tipos de programas de aplicación, existe la posibilidad de que si no se controla detalladamente el almacenamiento, se pueda originar un duplicado de información produciendo redundancia de los datos. Si la información esta contenida más de una vez en un dispositivo de almacenamiento, puede dar lugar a inconsistencia de los datos, esto es, que diversas copias de un mismo dato no concuerden entre sí. Por ejemplo, suponiendo que se tienen diferentes registros con la misma información y actualizamos solamente uno de ellos con una definición reciente del término *X*, los demás registros permanecerán con la definición anterior y esto provocará inconsistencia en el sistema.

3.2.2 Dificultad para tener acceso a los datos

Un sistema de base de datos debe contemplar una interfaz que le facilite al usuario el manejo de los datos, así como cumplir con sus requerimientos. Por ejemplo, supongamos que el usuario quiere saber la fuente de donde provino la definición del término *X*. Puesto que esta situación no fue prevista en el diseño del sistema, no existe ninguna aplicación de consulta que permita este tipo de solicitud, lo que ocasiona una deficiencia de dicho sistema.

3.2.3 Aislamiento de los datos

Puesto que los datos están repartidos en varios registros, si éstos no están relacionados unos con otros, ello puede causar que nunca se tenga acceso a cierta información por encontrarse aislada del sistema; esto va a generar que aunque se tenga el diseño apropiado no se obtengan los resultados esperados.

3.2.4 Anomalías del acceso concurrente

Para mejorar el funcionamiento global del sistema y obtener un tiempo de respuesta más rápido, muchos sistemas permiten que múltiples usuarios actualicen los datos simultáneamente. En un entorno así la interacción de actualizaciones concurrentes puede dar por resultado datos inconsistentes. Para prevenir esta situación, debe mantenerse la supervisión en el sistema por medio del DBA.

[2] http://atenea.udistrital.edu.co/profesores/jdimate/basedatos1/portada.htm

3.2.5 Problemas de seguridad

La información de toda base de datos es importante, por tal motivo se debe considerar el control de acceso a los mismos. No todos los usuarios pueden acceder a cierta información; para que un sistema de base de datos sea confiable debe mantener un grado de seguridad que garantice la autenticación y protección de los datos. Hay que proporcionarle a los usuarios finales una visión abstracta de los datos, esto se logra ocultando ciertos detalles de cómo se almacenan y mantienen los datos. Por ejemplo, no es conveniente que un usuario final tenga acceso a la información de administración del sistema, ya que puede modificarla o borrarla por error y generar problemas en el sistema. Por ello no se le deben presentar estas opciones.

3.2.6 Problemas de integridad

La integridad en las bases de datos depende de que: exista la menor redundancia posible y se elimine la inconsistencia de los datos; se tenga previsto dentro del diseño todas las posibles consultas que podría hacer el usuario al banco terminológico, de forma que no encuentre obstáculo alguno para acceder a la información; el DBA mantenga una buena administración del sistema para controlar el acceso concurrente de los usuarios; se controle al máximo la seguridad en el sistema para que los usuarios no provoquen pérdida de información del banco; se garantice que los usuarios no podrán hacer modificación alguna sin la debida autorización.

3.3 Análisis

En esta sección haremos un análisis del sistema que nos servirá para apoyar los requerimientos que se delimitan en el Capítulo 2.

A continuación, se desglosan tres objetivos fundamentales que debe cumplir un sistema en su etapa de análisis[3], con el fin de delimitar y visualizar los alcances del banco terminológico.

3.3.1 Identificación de las necesidades del cliente

Con base en los requerimientos que se analizaron en el Capítulo 2, el GIL tiene la necesidad de un banco terminológico con las siguientes características:

- Es necesaria una herramienta que apoye a las diferentes líneas de investigación que se están desarrollando en el GIL.
- Se requiere un banco terminológico para almacenar en un solo lugar toda la información compilada por el GIL, con el fin de que los usuarios puedan obtener una respuesta a sus peticiones de información de una forma agradable, rápida y segura, desde cualquier lugar en el que se encuentren.

[3] Pressman (1999)

- Se necesita que el banco resida en un sitio Web, con el fin de que los usuarios puedan acceder a él vía Internet.

3.3.2 Viabilidad del Sistema

El análisis de viabilidad del sistema tiene cuatro áreas por desarrollar:

- Viabilidad económica. Consiste en una evaluación del costo del desarrollo frente al beneficio final producido por el sistema desarrollado.

- Viabilidad técnica. Se trata de un estudio de la funcionalidad, del rendimiento y de las restricciones que pueden afectar a la posibilidad de realización de un sistema.

- Viabilidad legal. Determina cualquier infracción, violación o ilegalidad que pudiera resultar del desarrollo del sistema.

- Alternativas. Evalúa los enfoques alternativos para el desarrollo del sistema.

Sin embargo, durante la evolución de esta investigación se realizó un estudio a conciencia sobre este punto y se llegó a la conclusión de que los costos, el software, las cosas legales, etc. se toman por dados, debido a que el banco es una herramienta que se elaborará con el fin de apoyar al GIL en sus diferentes líneas de investigación.

3.3.3 Asignación de Funciones

El banco terminológico contará con una interfaz Web, que conlleve tres diferentes roles: administrador del banco terminológico, usuario del banco y visitante. Cabe denotar que el DBA es un administrador que tiene el control total del sistema y es el que asignará los permisos a las bases, dará de alta y baja a los usuarios de las mismas y asignará los roles.

- Administrador del banco terminológico (DBO, Data Base Owner). Esta interfaz es la principal de todas, ya que con ella se puede hacer todo tipo de operaciones en el banco terminológico. En ella, el DBO funge como propietario de una, varias o todas las bases de datos que contiene el banco. Las operaciones que puede realizar con esta interfaz son consultar toda la información (sin ninguna restricción) que se encuentra almacenada en el banco, así como insertarla, borrarla y/o modificarla.

- Usuario del banco. La interfaz para el usuario le permitirá consultar cierta información del banco (este tipo de usuario no tiene necesidad de ver la información de la administración del sistema), así como insertarla y modificarla sin poder borrarla. El usuario del banco no tiene permitido eliminar información e insertar nuevas fuentes en el banco terminológico, ya que esta acción tiene que estar autorizada y avalada por los expertos. En caso de que tenga la necesidad de borrar cierto contenido del banco, hará la petición de eliminación al sistema y

automáticamente le será enviado un correo al DBO con su petición sin borrar la información.

- Visitante: Para el caso de los visitantes, se presentará una interfaz para la consulta básica de un diccionario de terminología especializada.

El banco terminológico contará con un enlace para contactar directamente al administrador del sistema, con el fin de que todos los usuarios, sea cual sea su función, puedan enviar sus quejas o sugerencias con relación al mismo.

3.4 Diseño

En esta sección hablaremos del software que se propuso para el desarrollo del sistema; en este sentido mostraremos el diagrama jerárquico funcional del banco terminológico, el diagrama conceptual, el modelo físico y lógico. La creación de las bases de datos son parte del Capítulo 4, por lo que no se tratará el tema en el presente capítulo.

3.4.1 Propuesta de software

En este punto vamos a delimitar la plataforma que se va a utilizar en el desarrollo del banco terminológico así como el DBMS, el lenguaje de programación utilizado y el servidor Web. Así, debido a los requerimientos del GIL se decidió utilizar los siguientes recursos de software:

- El sistema operativo que se utilizará como plataforma será Windows 2000 Advanced Server®[4], ya que incluye mejoras para los servicios de red y aplicaciones. Suministra una mayor confiabilidad y escalabilidad, reduce los costos de computación mediante servicios de administración eficaz y flexible. Proporciona una base óptima para ejecutar aplicaciones en Internet, además de que es totalmente compatible con el manejador de bases de datos SQL Server 2000.

- Los lenguajes de programación que se utilizarán serán:

  a) PERL (Practical Extraction and Report Language). Es un lenguaje de programación medianamente nuevo, el cual surgió de algunas herramientas de UNIX. Sirve más que nada para labores de procesamiento de texto y últimamente se ha consolidado como lenguaje de programación en aplicaciones para Internet[5].

  b) HTML (HyperText Markup Language). Es una colección de estilos (indicados por etiquetas) que definen los distintos componentes de los documentos de World Wide Web[6].

[4] http://www.microsoft.com/windows2000/es/advanced/help/
[5] http://www.geocities.com/SiliconValley/Station/8266/perl/
[6] http://docs.indymedia.org/view/Local/HTMLGidaGazteleraz#A1.1

c) JavaScript. Es un lenguaje de programación muy fácil de usar que surgió a partir de Java y se pueden escribir scripts que funcionan en el entorno de una página Web, interpretado por un explorador[7].

- Como DBMS utilizaremos SQL Server 2000®[8], debido a que se trata de una completa solución de base de datos y análisis, la cual proporciona el rendimiento, la escalabilidad y la confiabilidad que requieren los exigentes entornos Web. La nueva compatibilidad con el Lenguaje de Marcado Extensible (XML, Extensible Markup Language) y el Protocolo de Transferencia de Hipertexto (HTTP, Hypertext Transfer Protocol) simplifican el acceso a los datos y su intercambio, al tiempo que las capacidades eficaces de análisis optimizan el valor de los datos. Las características de disponibilidad mejoradas maximizan el tiempo de actividad, en tanto las funciones de administración avanzadas automatizan las tareas rutinarias y las herramientas de programación y los servicios mejorados aceleran el desarrollo.

- Por último, como servidor Web utilizaremos Internet Information Server (IIS) 5.0®[9] Las características de IIS permiten compartir fácilmente documentos e información a través de Internet. Además, es posible distribuir aplicaciones escalables y confiables basadas en Internet e incorporar aplicaciones y datos existentes al Web.

3.4.2 Diagrama jerárquico funcional del sistema

Con el diagrama jerárquico funcional del sistema se mostrarán los módulos principales que constituyen el banco terminológico, así como las relaciones que tienen entre todos ellos (Figura 3-1).

Como primera interfaz, el usuario visualizará una ventana de login y password con lo cual el sistema autentificará el rol del usuario: según los permisos que se le hayan asignado, será el rol que tendrá. Como se mencionó anteriormente (3.3.3), hay tres roles en el banco. Inmediatamente se le presenta una interfaz que indica al usuario cuál base de conocimientos es sobre la que desea trabajar.

El banco terminológico le permitirá al "DBO" hacer cualquier tipo de operación sobre el sistema (consultar, insertar, modificar y/o eliminar), y estas opciones serán aplicadas en las diferentes opciones de búsqueda con las que cuenta el banco (términos, definiciones, fuentes, sinónimos, área temática y equivalencias). Hay tres escenarios en los que no se hace ninguna operación (ver completo, lista de caracteres especiales y ayuda): la primera sirve para visualizar las definiciones en pantalla completa, esto con el fin de que el usuario pueda imprimir y tener una mejor perspectiva de la información que se le presenta; en la segunda le muestra al usuario la forma de insertar caracteres especiales

[7] http://www.ciudadfutura.com/javascriptdesdecero/intro2.htm
[8] http://www.microsoft.com/latam/sql/evaluation/overview/2000/datasheet.asp
[9] http://www.microsoft.com/windows2000/es/advanced/help/

dentro del banco terminológico, tales como: α, β, μ, etc.; y en la tercera es la ayuda que tiene el usuario para que sepa cómo navegar dentro del banco terminológico.

Por otro lado, para el escenario que presenta el sistema al "usuario del banco", éste le permite realizar consultas, insertar, modificar y hacer una petición de eliminación de información en caso que así lo requiera. Estas operaciones las puede realizar en las diferentes opciones de búsqueda, con excepción de la opción de fuentes, donde únicamente puede hacer una operación de consulta por los motivos ya mencionados (3.3.3).

Por último el escenario de "visitante", le permite al usuario hacer cualquier consulta de información en todas las opciones del sistema.

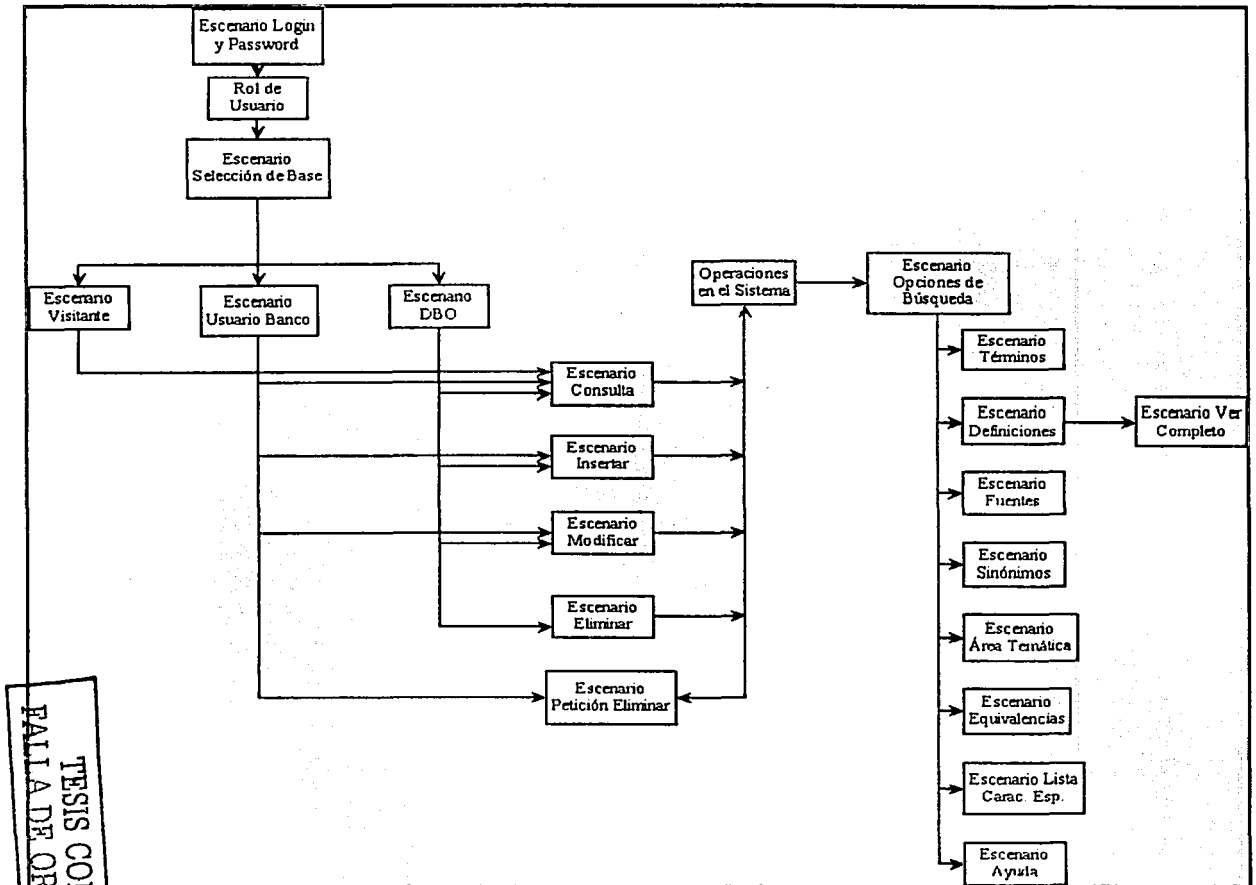


Figura 3-1. Diagrama jerárquico funcional

3.4.3 Diseño de las bases de datos

Las bases de datos que se diseñarán permitirán a los usuarios manipular toda la información concentrada en un solo lugar vía Internet, sin olvidar el rol que tendrá cada uno de ellos.

Dado que la información que contiene el GIL se encuentra en bases de datos aisladas (2.2.2), el diseño partió de éstas para unificar toda la información. Considerando la migración de los datos, se respetó la estructura con la que contaban. El diagrama entidad-relación de las bases de datos originales se muestran a continuación (Figura 3-2).

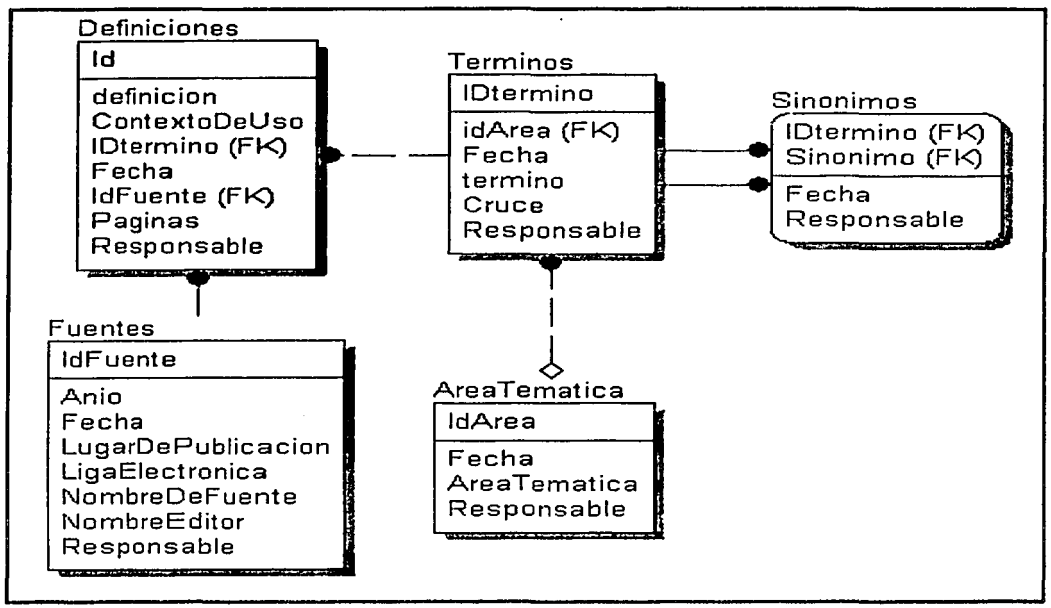


Figura 3-2. Diagrama original de las bases de datos

3.4.3.1 Definición de entidades

Tomando como punto de partida las bases de datos originales (Figura 3-2), fue necesario rediseñarlas, así como adicionar nuevas entidades para crear el banco terminológico. En la tabla 3-1 se muestran las entidades que resultaron después de adicionar las nuevas y rediseñar las que ya se tenían, así como sus respectivos atributos, llaves primarias (PK) y llaves foráneas (FK) que contendrán cada entidad:

Tabla 3-1. Entidades resultantes para el banco terminológico

Entidad	Atributos
BASE	IDBASE (PK) IDIDIOMA (FK) BASE PÚBLICA RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN
EQUIVALENCIA	IDEQUIVALENCIA (PK) IDTÉRMINO (FK) INGLÉS FRANCÉS ALEMÁN PORTUGUÉS RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN
TERMINO	IDTÉRMINO (PK) IDBASE (FK) TÉRMINO RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN
DEFINICIÓN	IDDEFINICIÓN (PK) IDTÉRMINO (FK) IDFUENTE (FK) DEFINICIÓN CONTEXTO DE USO PÁGINAS RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN
ÁREA TEMÁTICA	IDÁREA (PK) IDTÉRMINO (FK) IDBASE (FK) ÁREA TEMÁTICA RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN
SINÓNIMO	SINÓNIMO (PK) IDTÉRMINO (FK) RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN

FUENTE	IDFUENTE (PK) IDBASE (FK) ANIO LUGAR DE PUBLICACION LIGA ELECTRONICA NOMBRE DE FUENTE NOMBRE EDITOR EDITORIAL AUTOR IDENTIFICADOR RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN

El diagrama entidad-relación que resultó después de haber definido nuestras entidades se muestra en la figura 3-3. Sin embargo, como se puede observar en la figura, resultó que la entidad de área temática tiene una relación de muchos a muchos con la entidad de términos y la entidad base. Utilizamos la herramienta de Power Designer® para romper estas relaciones y cumplir con las reglas de normalización en las bases de datos. El resultado que obtuvimos con esta herramienta fueron 2 entidades adicionales, las cuales nos aseguran que las bases de datos cumplen con las cuatro reglas de normalización. Las nuevas entidades se ilustran en la tabla 3-2.

Tabla 3-2. Entidades nuevas

Entidad	Atributos
ÁREA-BASE	IDBASE (FK) IDÁREA (FK) RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN
ÁREA-TÉRMINO	IDTÉRMINO (FK) IDÁREA (FK) RESPONSABLE FECHA RESPONSABLE MODIFICACIÓN FECHA MODIFICACIÓN

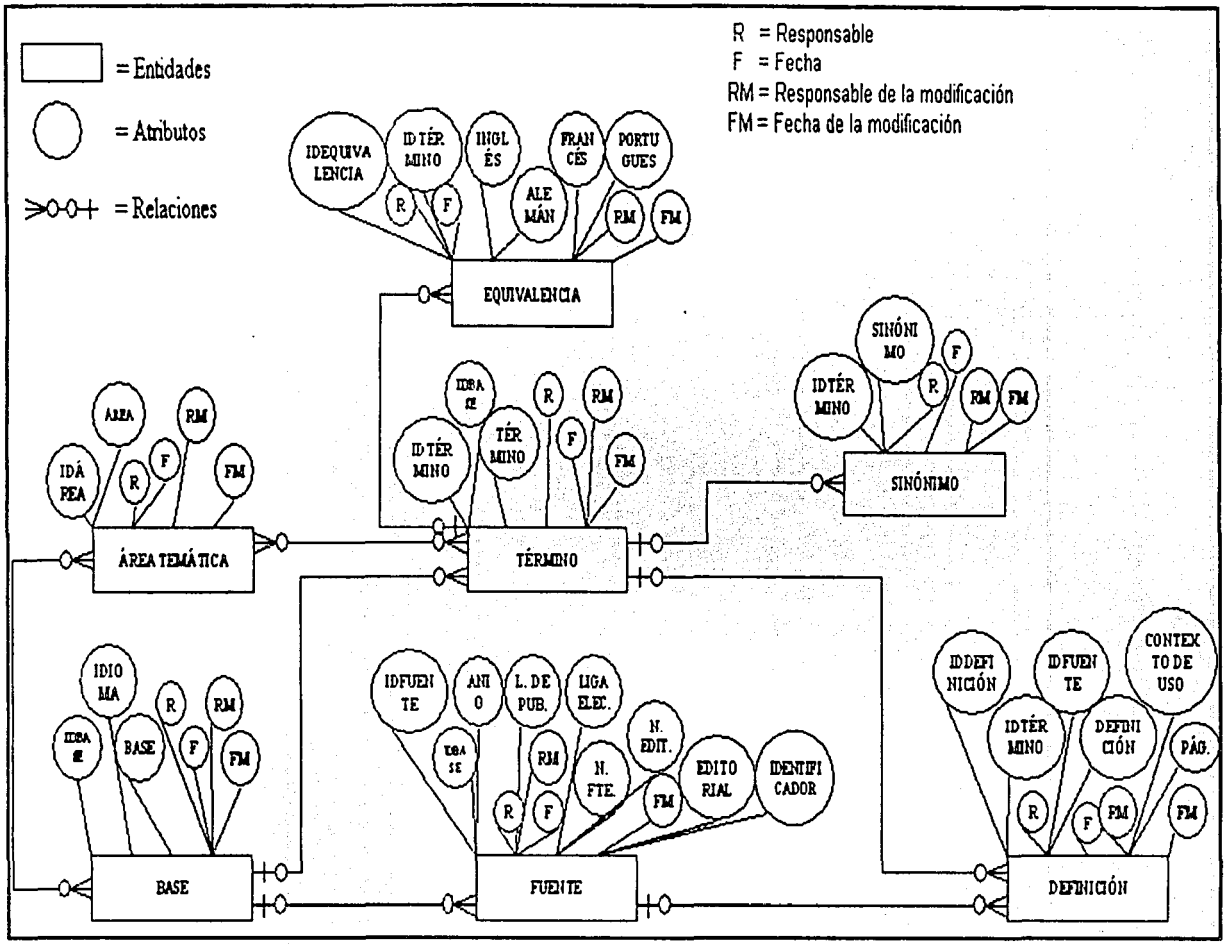


Figura 3-3. Entidad-relación del banco terminológico

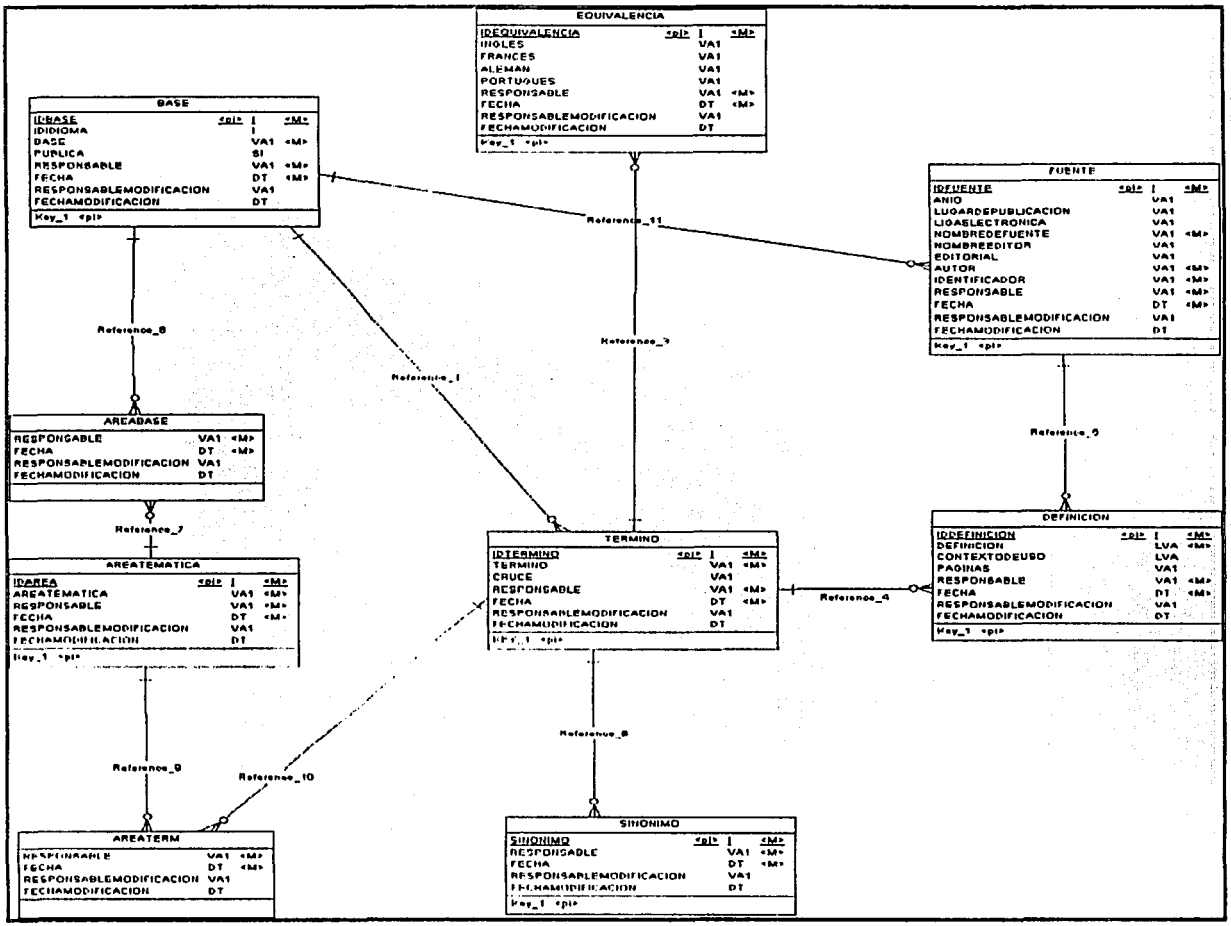


Figura 3-4. Modelo conceptual del banco terminológico

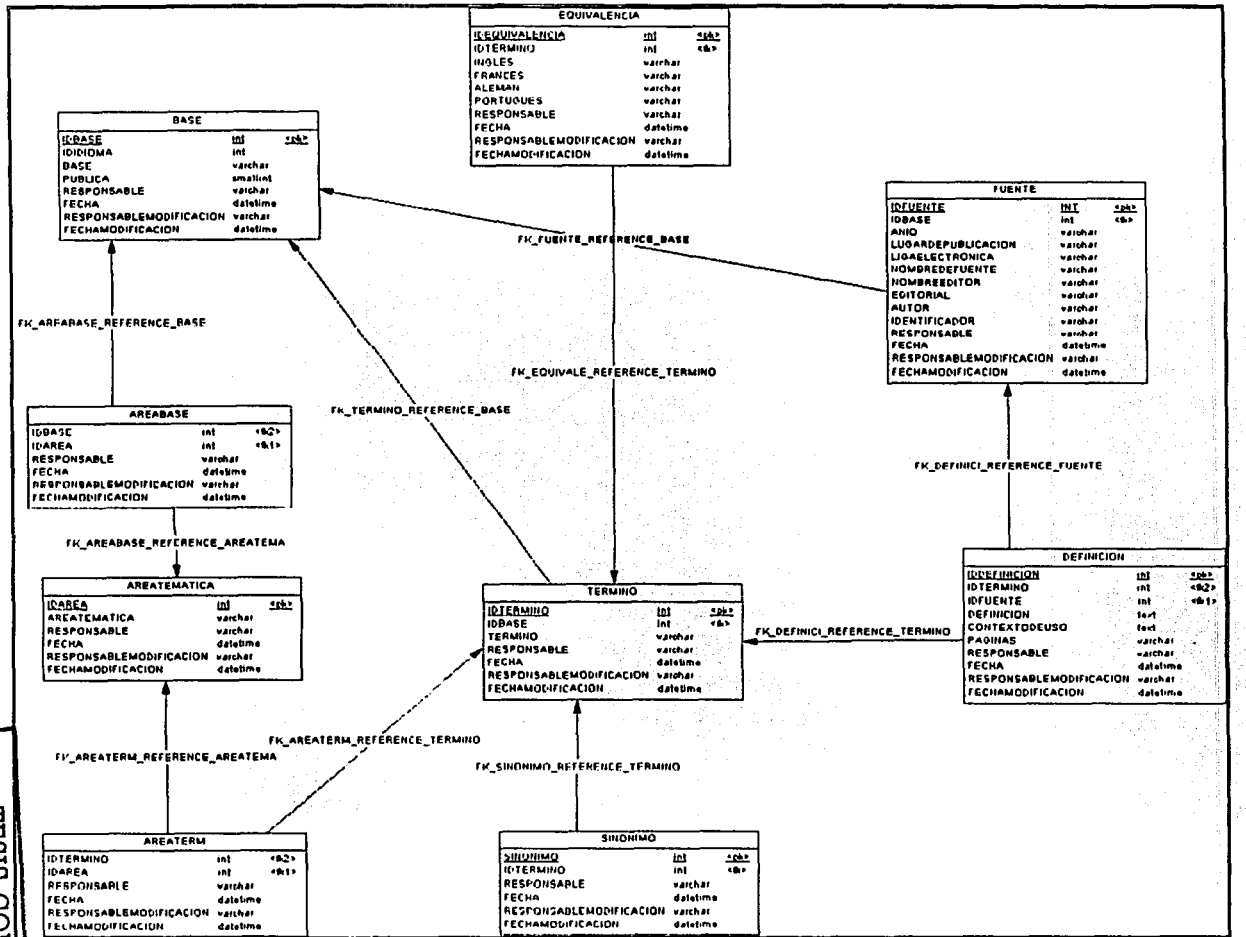


Figura 3-5. Modelo físico del banco terminológico

En la figura 3-4 se representa el modelo conceptual de la base de datos del banco terminológico, el cual nos muestra las relaciones existentes entre todas las tablas; y en la figura 3-5 se muestra el modelo físico de las bases, que nos representa los tipos de datos que contienen las tablas así como sus llaves primarias y foráneas. Se puede observar que ya no existen las relaciones muchos a muchos que encontramos en el modelo entidad relación (Figura 3-3). Éstas fueron eliminadas gracias a la herramienta Power Designer® que nos verifica y corrige este tipo de relaciones.

3.4.3.2 Diccionario de datos

El diccionario de datos contiene las características de las entidades y atributos que definen la estructura de las bases de datos del sistema. El objetivo primordial del diccionario de datos es facilitar el control de cada una de las entidades, atributos que forman parte de la base, los tipos de datos, las llaves primarias y foráneas que maneja cada tabla así como los datos que pueden o no ser ingresados por el usuario (NULL). La tabla 3-3 nos muestra el diccionario de datos del banco terminológico.

Tabla 3-3. Diccionario de datos del banco terminológico

Nombre de la Base	Descripción de la Base	Nombre del Atributo	Descripción del Atributo	Tipo de Dato	PK	FK	NULL
BASE	Contiene las bases existentes	IDBASE	Identificador de la base	Int	◄		
		IDIDIOMA	Identificador del idioma	Int		◄	
		BASE	Nombre de la base	Varchar (20)			
		PÚBLICA	Indica si es pública o no la base	Smallint			◄
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄
EQUIVALENCIA	Contiene las equivalencias de los términos en los diferentes idiomas que maneja el banco	IDEQUIVALENCIA	Identificador de la equivalencia	Int	◄		
		IDTÉRMINO	Identificador del término al que se está haciendo referencia	Int		◄	
		INGLÉS	Equivalencia en inglés	Varchar (20)			◄
		FRANCÉS	Equivalencia en francés	Varchar (20)			◄
		ALEMÁN	Equivalencia en alemán	Varchar (20)			◄
		PORTUGUÉS	Equivalencia en portugués	Varchar (20)			◄
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que la modificó	Datetime			◄
TERMINO	Contiene toda la terminología	IDTÉRMINO	Identificador del término	Int	◄		
		IDBASE	Identificador de la base a la que se está haciendo referencia	Int		◄	

		TÉRMINO	Nombre del término	Varchar (20)			
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄
DEFINICIÓN	Contiene las definiciones de los términos	IDDEFINICIÓN	Identificador de la definición	Int	◄		
		IDTÉRMINO	Identificador del término al que se está haciendo referencia	Int		◄	
		IDFUENTE	Identificador de la fuente a la que se está haciendo referencia	Int		◄	
		DEFINICIÓN	Nombre de la definición	Text			
		CONTEXTO DE USO	Contexto definitorio del término y la definición a los que se está haciendo referencia	Text			◄
		PÁGINAS	Páginas de la fuente de donde proviene la información	Varchar (20)			◄
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄
ÁREA TEMÁTICA	Contiene las áreas temáticas de las diferentes bases	IDÁREA	Identificador del área temática	Int	◄		
		ÁREATEMÁTICA	Nombre del área temática	Varchar (20)			
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE	Fecha en que se modificó	Datetime			◄

		MODIFICACIÓN					
SINÓNIMO	Contiene los sinónimos de los términos	SINÓNIMO	Identificador del sinónimo	Int	◄		
		IDTÉRMINO	Identificador del término al que se está haciendo referencia	Int		◄	
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄
FUENTE	Contiene las fuentes de donde fue extraída la información	IDFUENTE	Identificador de la fuente	Int	◄		
		IDBASE	Identificador de la base a la que se está haciendo referencia	Int		◄	
		ANIO	Año de la publicación de la fuente	Varchar (10)			◄
		LUGAR DE PUBLICACION	Lugar de la publicación	Varchar (50)			◄
		LIGA ELECTRONICA	Liga electrónica de referencia	Varchar (50)			◄
		NOMBRE DE FUENTE	Nombre de la fuente	Varchar (50)			
		NOMBRE EDITOR	Nombre del editor	Varchar (50)			◄
		EDITORIAL	Nombre de la editorial	Varchar (50)			
		AUTOR	Nombre del autor de la fuente	Varchar (50)			
		IDENTIFICADOR	Nombre en formato corto de la fuente para mostrarlo en la interfaz	Varchar (50)			
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄
ÁREA-BASE	Relación que sirve de conexión entre la	IDBASE	Identificador de la base a la que se está haciendo referencia	Int		◄	
		IDÁREA	Identificador del área temática a la	Int		◄	

	entidad de área temática y base		que se está haciendo referencia				
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄
ÁREA-TÉRMINO	Relación que sirve de conexión entre la entidad de área temática y término	IDTÉRMINO	Identificador del término al que se está haciendo referencia	Int		◄	
		IDÁREA	Identificador del área temática a la que se está haciendo referencia	Int		◄	
		RESPONSABLE	Responsable de la inserción de la información	Varchar (20)			
		FECHA	Fecha en que se insertó	Datetime			
		RESPONSABLE DE MODIFICACIÓN	Responsable de la modificación de la información	Varchar (20)			◄
		FECHA DE MODIFICACIÓN	Fecha en que se modificó	Datetime			◄

Capítulo 4. Implantación del banco terminológico

Esta etapa se centrará en la forma en como será implantado el banco terminológico para el Grupo de Ingeniería Lingüística. Se describirá la plataforma para la implantación y los lenguajes de programación usados.

Se presentarán las interfaces usuario-máquina y la serie de pasos que se generaron como preámbulo, para obtener el sistema final.

4.1 Integración del sistema

Para obtener el banco terminológico al cual tendrán acceso los usuarios, se procederá a integrar las bases de datos que conforman el banco, así como implantar las interfaces con los lenguajes de programación PERL, HTML y Java Script, y a elaborar los programas (motor de búsqueda) que funcionarán de enlace entre las bases y las interfaces usuario-máquina.

Los motivos de utilizar el software antes mencionado radica en:

- En principio PERL (Practical Extraction and Report Lenguaje) es un software libre que es del tipo interpretativo, esto significa que al escribir el código, el interprete lo descifra y ejecuta las instrucciones, esto evita la necesidad de compilar el programa y por consiguiente el tiempo de ejecución es menor. Por otro lado es un lenguaje que está diseñado para trabajar con cadenas de texto y dado que el banco terminológico en esencia es texto, PERL nos será de gran utilidad.

- En el caso de HTML (Hyper Text Markup Language) es la forma estándar de describir los contenidos y la apariencia de las páginas en el World Wide Web. HTML se compone de atributos y valores incluidos entre pares de etiquetas las cuales describen cada elemento de una página Web, como por ejemplo un párrafo de un texto, una tabla o una imagen. Dado que se requiere que el banco se encuentre en Internet y además resaltar ciertas partes de texto es necesario contar con HTML.

- Finalmente para el caso de Java Script, durante el desarrollo del banco se necesitan scripts que trabajen de forma dinámica y éste lenguaje de programación cumple con las expectativas que se solicitan, además de ser fácil de usar, compatible con HTML y PERL.

4.1.1 Especificaciones del hardware

Hay que tener presente las necesidades que se tienen para la instalación del software utilizado en la elaboración del banco terminológico. Por ello, se procederá a hacer un análisis de los distintos tipos de software que se utilizarán en la implantación del sistema, e igualmente saber cuáles son sus requerimientos mínimos de hardware (Tabla 4-1).

Tabla 4-1. Requerimientos mínimos de hardware para los diferentes tipos de software

Software / Hardware	Microsoft Windows 2000 Server Advanced	Microsoft SQL Server 2000	IIS	PERL	HTML	Java Script
Procesador	Uno o más a 133 MHz o más rápidos	166 MHz	486 a 33 MHz o superior	166 MHz	486 a 33 MHz o superior	133 MHz
Memoria RAM[1]	64 Mb	64 Mb	16 MB	16 MB	16 MB	16 MB
Espacio libre en disco duro	Partición de 850 Mb con 650 Mb	180 MB	40 MB	52 MB	---	---
Sistema Operativo	---	Windows 2000 Advanced Server	Windows 95	Windows 98	Windows 95	Windows 95
Navegador	---	---	---	---	Microsoft Internet Explorer 3.x	Microsoft Internet Explorer 3.x

De acuerdo con el análisis anterior, se llegó a la conclusión de utilizar una computadora con las especificaciones de hardware señaladas en la tabla 4-2.

Tabla 4-2. Especificaciones de hardware

Dispositivo	Características
Sistema Operativo	Microsoft Windows 2000 Advanced Server
Versión	5.0.2195 Service Pack 3 Build 2195
Fabricante del Sistema Operativo	Microsoft Corporation
Fabricante de la PC	Hewlett-Packard
Modelo de la PC	HP Brio
Procesador	Genuine Intel 551 Mhz
Versión del BIOS	Award Modular BIOS v6.00PG
Windows Directory	C:\WINNT
Memoria física total (RAM)	261,616 KB
Capacidad del disco duro	40 GB
Navegador	Microsoft Internet Explorer V.6.0

4.1.2 Creación de las bases del banco terminológico

El query que se utilizará para la elaboración de las bases de datos del banco terminológico será creado automáticamente a partir del modelo físico que se diseñó con la herramienta de Power Designer (3.4.3.1). Sin embargo, se le tendrán que hacer modificaciones al query para adaptarlo al DBMS SQL Server 2000.

[1] RAM (Random Acces Memory): Memoria de Acceso Aleatorio

Una vez adaptado el query, se ejecutará directamente en el DBMS y se crearán de forma automática las bases de datos que conforman el banco terminológico.

Enseguida se muestra parte del código que se generó con Power Designer y que posteriormente se adaptó a SQL Server 2000:

```
/*==============================================================*/
/* Nombre de la base de datos: BANCO TERMINOLÓGICO */
/* Propietario de la bases de datos: Gabriel Garduño */
/* Nombre del DBMS:     SQL SERVER 2000                */
/* Creado en:    09/01/2003 04:00:28 p.m.              */
/*==============================================================*/

/*==============================================================*/
/* Crear la base de datos: bancoTerminologico                   */
/*==============================================================*/
create database  bancoTerminologico;      /* Instrucción que crea la base de datos "bancoTerminologico */

/*==============================================================*/
/* Creación de la tabla BASE                                    */
/*==============================================================*/

drop table if exists BASE;                /* Instrucción que borra la tabla "BASE" si es que existe */
create table BASE          (              /* Inicio de la instrucción para crear la tabla "BASE" */
   idBase          int        not null,   /* Campo "idBase" de tipo entero y no acepta nulos */
   idIdioma        int        null,       /* Campo "idIdioma" de tipo entero y acepta nulos */
   base   varchar  not null,              /* Campo "base"de tipo carácter variable y no acepta nulos */
   publica         smallint null,         /* Campo "publica" de tipo entero reducido y no nulos */
   responsable     varchar  not null,     /* Campo "responsable"de tipo carácter variable y no acepta */
                                          /* nulos */
   fecha           datetime not null,     /* Campo "fecha" de tipo fecha y no acepta nulos */
   responsableModificacion          varchar null,    /* Campo "responsableModificacion" de tipo carácter */
                                                     /* variable y acepta nulos */
   fechaModificacion       datetimenull   /* Campo "fechaModificacion" de tipo fecha y acepta nulos */
   primary key (idBase)                   /* Instrucción que inserta la llave primaria "idBase" */
):                                        /* Fin de la instrucción que crea la tabla */

/*==============================================================*/
/* Creación de llaves foráneas                                  */
/*==============================================================*/

alter table AREABASE                      /* Instrucción que altera los datos de la tabla "AREABASE" */
add foreign key FK_AREABASE_REFERENCE_BASE (idBase)     /* Instrucción que nos indica que el
                                          /* campo "idBase" será la llave foránea con el nombre FK_A... */
references BASE (idBase)          /* Instrucción que nos indica que el campo "idBase" proviene de */
                                  /* de la tabla "BASE" */
```

La documentación completa del código para generar las bases de datos del banco terminológico, se encuentra disponible para su consulta en el Grupo de Ingeniería Lingüística, del Instituto de Ingeniería, UNAM.

4.1.3 Seguridad del sistema

Es vital tomar en cuenta la seguridad que rodea al banco terminológico, dado que la información que contiene es muy importante; por ello se mantendrá un estricto control en el acceso de usuarios a la información, de tal forma que quienes no tengan permisos para agregar, modificar y/o eliminar la información, no lo puedan hacer de ninguna forma. Del mismo modo, hay que proteger los datos de las personas que son ajenas al uso del banco y dañen la información. Para este caso, nos aseguramos que el sistema operativo y el DBMS proporcionarán una gran confiabilidad en su uso y asegurarán que el sistema estará protegido todo el tiempo.

Haciendo un minucioso estudio del software utilizado, confirmamos su uso por las características más importantes que encontramos en ellos y las cuales mostramos enseguida:

- Sistema operativo Windows 2000 Advanced Server ®[2]

  a) El IIS Lockdown Tool permite a los usuarios bloquear el sistema y esto es suficiente para proteger la PC contra vulnerabilidades desconocidas de IIS security (por ejemplo el virus "código rojo"), incluso si los parches para estas debilidades del sistema no han sido instalados.

  b) URLScan ayuda a proteger al servidor web asegurándose de que sólo responderá a peticiones legítimas. Muchos ataques contra el servidor web incluyen peticiones que son poco habituales, pueden ser demasiado largas o pueden incluir combinaciones de caracteres muy difíciles de detectar. URLScan filtra estas peticiones y se asegura que nunca lleguen al servidor y por lo tanto no tengan éxito.

  c) HFNetChk es una herramienta que da al administrador la posibilidad de verificar el estado de los parches de la máquina. La herramienta hace referencia a una base de datos XML (eXtended Markup Language) que está constantemente actualizada por Microsoft®. HFNetChk escaneará tanto el sistema remoto como el local cuidando los parches.

  d) Microsoft Personal Security Advisor (MPSA)® es una aplicación web que facilita información completa y precisa, mediante un informe, sobre la seguridad del sistema y las recomendaciones para mejorarlo, tales como parches perdidos, contraseñas de fácil acceso, escenarios de seguridad de Internet Explorer y Outlook Express, etc.

- DBMS Microsoft SQL Server 2000®[3]

  Las dos categorías de suma importancia que nos brinda SQL Server 2000 son:

[2] http://www.microsoft.com/spain/servidores/windows2000/seguridad/iis.asp
[3] http://www.microsoft.com/latam/technet/articulos/sql/operation_guide/art03/default.asp

a) *Directivas de seguridad.* Son un conjunto de protocolos y procedimientos para implantar la seguridad en un entorno de tecnología de la información. El principal énfasis de la seguridad que nos presenta SQL Server 2000 en esta categoría, es sobre la protección de datos durante el almacenamiento (como en bases de datos, copias de seguridad y archivos) o durante la transferencia (como en consultas, DTS[4] y duplicaciones). Además cuenta con controles que sirven para proteger, detectar y corregir "puertas abiertas", lo que niega o evita todo tipo de acceso no autorizado a las bases de datos.

b) *Autenticación.* SQL Server 2000 cuenta con dos formas de autenticación en el acceso de los usuarios: modo de autenticación de Windows (el modo predeterminado) o el modo mixto (tanto con autenticación de Windows como con autenticación de SQL Server). Estas características son propias de las directivas de seguridad de Windows 2000, pero también se encuentran implantadas en SQL Server. Trabajando en conjunto el sistema operativo y el DBMS, ambos controlan la caducidad y los atributos de las contraseñas, las audiciones, los bloqueos de cuentas y la autenticación mutua con Kerberos.

4.1.4 Migración de la información

Partiendo de que la información que pertenecerá al banco terminológico se encuentra en diccionarios especializados aislados (2.2.2), comenzamos a migrar la información de Microsoft SQL Server 7.0®, el cual era el DBMS en donde radicaban las bases de datos originalmente, a Microsoft Excel®.

Hicimos uso de la herramienta SQL Query Analizer® para extraer la información: para ejemplificarlo, mostramos uno de los datos con su formato original que contenía la tabla de términos, de la base de datos de Lingüística (Tabla 4-3).

Tabla 4-3. Información de la tabla de términos, de la base de datos de Lingüística

IdtérmIno	IdÁrea	Fecha	Término	Cruce	Responsable
1761	5	05/12/2001 01:12:21 p.m.	abecedario	NULL	RZacarias

Para modificar el formato que contenían las tablas originales, nos apoyamos en Excel para eliminar y adicionar los campos requeridos para el banco terminológico. Para ejemplificar el resultado de las modificaciones hechas a la información, e igualmente partiendo del ejemplo de la tabla 4-3, el formato final es el que se muestra en la tabla 4-4.

El campo "IdÁrea" que aparecía en la tabla 4-3, ya no existe en la tabla 4-4 porque al momento de diseñar las bases de datos, éste campo se eliminó de la tabla de términos y

[4] DTS (Data Transformation Services): Servicios de Transformación de Datos

aparece dentro de la relación que sirve de conexión entre la entidad del área temática y los términos (3.4.3.2).

Tabla 4-4. Información lista para migrarla al banco terminológico

Idtérmino	Idbase	Término
1761	2	abecedario

Responsable	Fecha	Responsable Modificación	Fecha Modificación
RZacarias	05/12/2001 01:12:21 p.m	NULL	NULL

Por último, para terminar con el proceso de migración, importamos toda la información desde Excel a SQL Server 2000 de forma automática.

4.2 Interfaces del banco terminológico y pruebas de funcionalidad

Tomando como referencia el diagrama jerárquico funcional (3.4.2), se generaron las diferentes interfaces del banco terminológico: DBO, usuario del banco y visitante (3.3.3). Los lenguajes de programación Perl, HTM y Java Script sirvieron de apoyo para la programación de las diferentes interfaces usuario-máquina. En total el banco terminológico consta de 39 archivos con una densidad de 4,000 líneas aproximadamente y un espacio en el disco duro de 296 KB.

Las pruebas de funcionalidad del sistema se presentarán en las propias imágenes de las interfaces, debido a que las búsquedas mostradas son ejemplos reales; por ello, los resultados que el sistema nos genera son verdaderos y confiables.

4.2.1 Acceso al banco terminológico

La entrada de los usuarios al banco, las bases a las que tiene acceso y las diferentes búsquedas, se muestran en las figuras 4-1, 4-2 y 4-3. No importa el rol que tengan, la interfaz será la misma para todos ellos.

En la figura 4-1, el usuario deberá introducir un login y un password, que previamente le serán asignados por el administrador del sistema, del modo que el banco terminológico detectará qué rol tiene y así presentarle la interfaz que le corresponde.

Cualquier usuario que desee consultar el banco terminológico lo podrá hacer en la página Web http://imsai.iingen.unam.mx/diccionarios/banco.htm, con el login: "visitante" y el password: "iling", que equivale al rol de "visitante" dentro del sistema. Con este acceso, los usuarios tienen la posibilidad de generar cualquier cantidad de consultas al banco y hacer uso de los diferentes tipos de búsqueda de información, de manera que puedan grabar o imprimir la información si así lo desean.

Esta interfaz se encuentra disponible en Windows Internet Explorer 6.0, para que en dado caso que el usuario no cuente con dicho sistema, lo pueda descargar automáticamente y visualice de manera precisa el banco en su PC. Además, existe un enlace para que los

usuarios puedan enviar vía correo electrónico sus dudas, comentarios o sugerencias al administrador del banco terminológico y administrador del sistema.

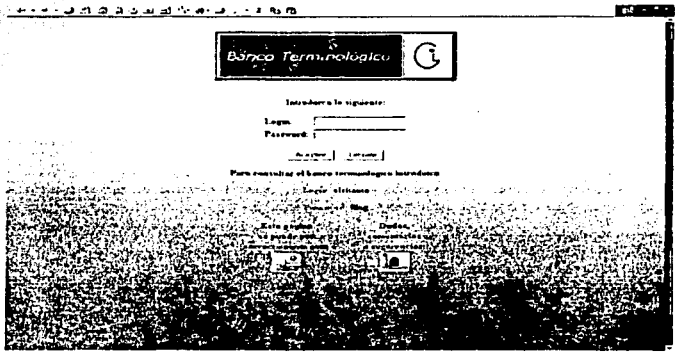


Figura 4-1. Login y password

En la interfaz que se presenta en la figura 4-2, se empleó el rol de "administrador del banco" porque nos muestra todas las bases de datos especializadas con las que cuenta el banco terminológico. Si el usuario tiene un rol de "usuario del banco", visualizará las bases en las que tenga permisos. Para el caso de los usuarios "visitantes", únicamente verán las bases que estén catalogadas como públicas.

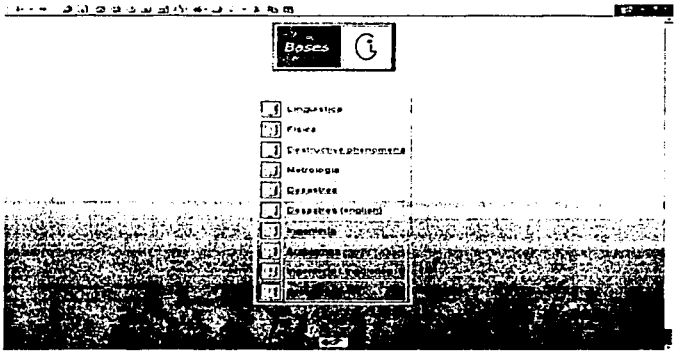


Figura 4-2. Bases disponibles al usuario

En la figura 4-3 se muestran las diferentes búsquedas de información posibles que pueden generar los usuarios dentro del banco terminológico. Además, el sistema proporciona una bienvenida que muestra la fecha, hora en la que accedió al banco y la base de datos especializada a la cual accedió el usuario.

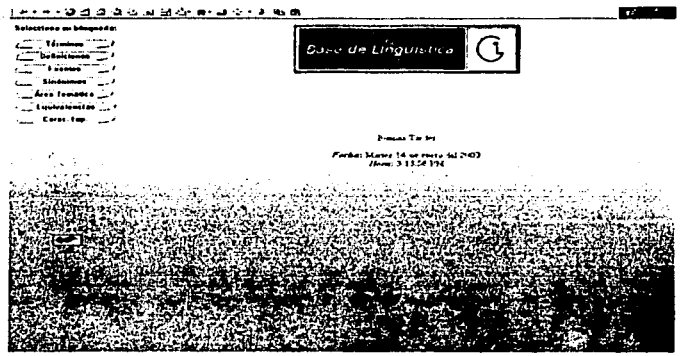


Figura 4-3. Búsquedas y bienvenida

Dentro de las búsquedas se encuentran datos asociados que varían dependiendo de lo que el usuario esté consultando. A continuación, se enumeran los resultados que obtendrán los usuarios al realizar alguna de las búsquedas con las que cuenta el banco terminológico:

- Términos: partimos del término y obtendremos su(s) área(s) temática(s) asociada(s) (pueden ser 1 o más a la vez), el responsable, la fecha, el responsable de modificación y la fecha de modificación. Se puede consultar más a detalle las funciones de la información mostrada en el diccionario de datos (3.4.3.2)

- Definiciones: partimos del término y obtendremos el área temática a la que pertenece, así como la(s) definición(es) almacenada(s) dentro del banco, su identificador asociado, el identificador de la fuente de donde proviene la información, los contextos definitorios que le pertenecen, las páginas de donde fue extraída la información, el responsable, la fecha, el responsable de modificación y la fecha de modificación.

- Fuentes: partiendo de un nombre en formato corto de la fuente, obtendremos el nombre completo de la fuente, su identificador asociado, el nombre completo de la fuente, autor de la fuente, el nombre del editor, la editorial, lugar de publicación, año de la publicación, liga electrónica, el responsable, la fecha, el responsable de modificación y la fecha de modificación.

- Sinónimos: partiendo del término, obtendremos el área temática a la que pertenece, el (los) sinónimo(s) que tenga dentro del banco terminológico, el responsable, la fecha, el responsable de modificación y la fecha de modificación.

- Área temática: partimos del área temática y obtendremos el responsable, la fecha, el responsable de modificación y la fecha de modificación.

- Equivalencias: nuevamente partimos del término y se obtendrá el área temática a la que pertenece, sus equivalencias en los idiomas inglés, francés y alemán, así como el responsable, la fecha, el responsable de modificación y la fecha de modificación.

4.2.2 Interfaz del DBO

Tomaremos como referencia al rol de administrador del banco (DBO), ya que es el más relevante dentro del sistema y los resultados que nos genera dentro de las búsquedas son los más completos. Cabe notar que este tipo de usuario puede agregar, modificar y/o eliminar la información, al contrario de los roles de usuario del banco y visitante.

Para realizar las pruebas de funcionalidad del sistema, tomaremos como base al término "abecedario" y la opción de búsqueda por "definiciones", por ser la más completa de todas las existentes. Se toma un caso real para que todas las interfaces aquí presentadas sean resultados verdaderos y confiables para el lector.

Dado que partimos del supuesto de que el banco terminológico se encuentra sin nada de información, lo primero que se hace es agregar una definición. La figura 4-4 nos presenta la interfaz que nos permite agregar una nueva definición al banco terminológico. Ésta muestra el término y el área temática a la que va a pertenecer la información y nos presenta estos datos fijos, de manera que el usuario no pierda de vista a quién va a pertenecer la información que introduzca. Resulta obligatorio introducir una definición y la fuente de donde proviene para que quede almacenada en el sistema: en el caso de que no se introduzca alguno de éstos datos, el sistema muestra una ventana de alerta, indicando que hace falta llenar algún registro para completar la operación. Los demás campos pueden o no ser introducidos.

Algo que es muy importante mencionar es la forma en que podemos introducir la información. Si el usuario va a introducir marcas tipográficas en el texto, el sistema detecta las marcas de HTML, de forma que puede introducir una definición o un contexto definitorio tal y como lo encontró en la fuente. Al momento de introducir el texto, el usuario tiene que colocar las etiquetas para **negrillas** xxx, *itálicas* xxx y/o subrayado <s>xxx</s>, envolviendo a las palabras que se quiere que mantengan el formato original.

Para ejemplificar esto, utilizaremos el contexto definitorio del término "abreviatura[5]":

> "Abreviación gráfica de una o varias palabras, que mantiene obligatoriamente el grafema inicial o éste y otro u otros más de la(s) palabra(s) que representa: don> D., doctor> Dr., señora> Sra., Puebla> Pue., administración> admon., Sociedad Anónima de capital variable> S.A. de C.V., San Luis Potosí> S.L.P., si Dios quiere>s.D.q., besa sus pies>b.s.p., descanse en paz>D.E.P. Cuando la abreviatura representa sintagmas o frases hechas suele denominarse también abreviatura compuesta."

[5] Diccionario básico

El texto se verá en la interfaz con el siguiente formato:

> "Abreviación gráfica de una o varias palabras, que mantiene obligatoriamente el grafema inicial o éste y otro u otros más de la(s) palabra(s) que representa: don> **D.**, doctor> **Dr.**, señora> **Sra.**, Puebla> **Pue.**, administración> **admon.**, Sociedad Anónima de capital variable> **S.A. de C.V.**, San Luis Potosí> **S.L.P.**, si Dios quiere>**s.D.q.**, besa sus pies>**b.s.p.**, descanse en paz>**D.E.P.** Cuando la abreviatura representa sintagmas o frases hechas suele denominarse también abreviatura compuesta".

Con este tipo de etiquetado estamos asegurando que se mantendrá la tipografía original de la fuente, tal y como la colocó el autor.

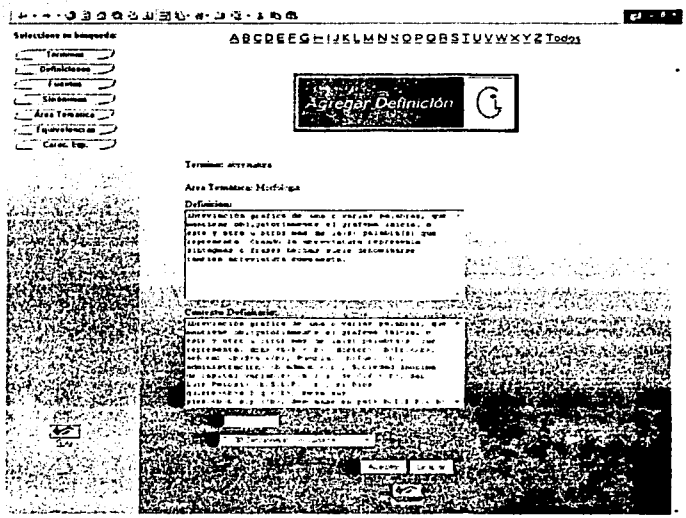


Figura 4-4. Agregar una definición

Cuando consultamos la información perteneciente a una definición (Figura 4-5), el sistema nos presenta todos los datos del término que seleccionamos: el término de referencia, el área temática a la que pertenece, el número de la definición dentro del sistema, la fuente de donde proviene, la definición, el contexto definitorio de donde fueron extraídos el término y la definición, el responsable que introdujo la definición y en qué fecha lo hizo. En caso de ser modificada la información, el sistema nos muestra el usuario que la modificó y en qué fecha. Finalmente, nos presenta las opciones de agregar una nueva definición al término mostrado, así como modificar o eliminar la información de la definición marcada.

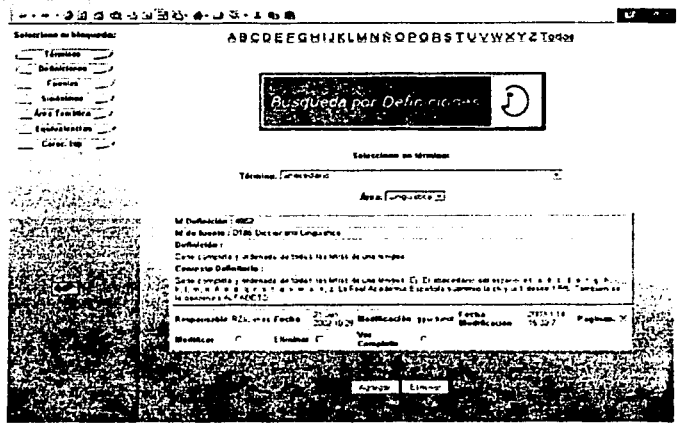


Figura 4-5. Consultando la definición de un término

El sistema tiene una opción para ver la definición, el contexto definitorio y la fuente del término que se esta consultando, en una pantalla completa en el caso de que el usuario desee imprimir o guardar la información que se le presenta (Figura 4-6).

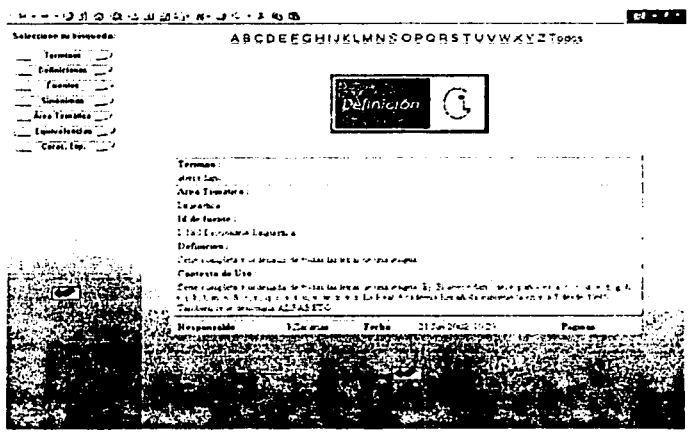


Figura 4-6. Ver la definición completa

En caso de que se desee modificar la información (Figura 4-7), el banco nos presenta toda la información que fue ingresada al momento de agregar la definición, correspondiente al término "abecedario" y dispuesta a ser modificada. Cabe resaltar que el sistema indica la fuente que se almacenó desde un principio y nos da la opción de modificarla si es que así se desea; en caso contrario dejará la fuente existente hasta el momento.

El sistema automáticamente guarda el login del usuario que modificó los datos y la fecha en que lo hizo; el usuario que ingresó por primera vez la información, así como la fecha en que lo hizo, quedan almacenados y estos datos no podrán ser modificados y/o eliminados en ningún momento, ya que son parte de la administración de los usuarios. Hay que señalar además que estos datos los emplea el GIL para recabar información estadística, la cual será utilizada para hacer conteos de los datos introducidos por cada usuario, así como las fechas en que lo hicieron.

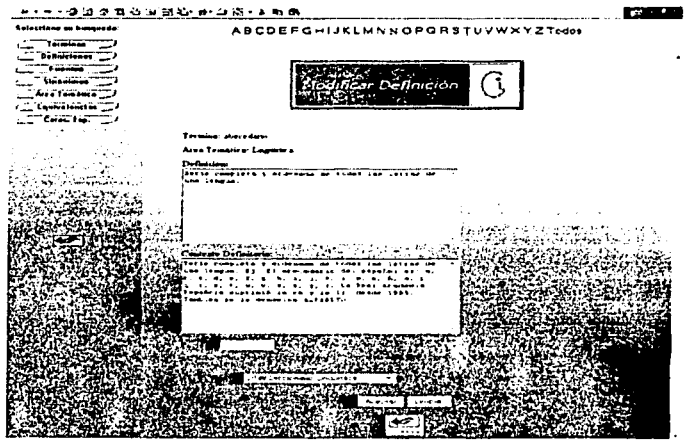


Figura 4-7. Modificar una definición

Cuando un usuario desea eliminar una definición, se debe estar consciente de que la información, una vez eliminada, no la podrá recuperar de ninguna forma. Dado que la información del banco terminológico es de suma importancia, se implantaron dos sistemas de protección de la información, de tal forma que si el usuario pulsa por error el botón de "Eliminar", el sistema le presentará una pantalla con un mensaje de alerta indicándole que no eligió ninguna definición para eliminar (Figura 4-8)

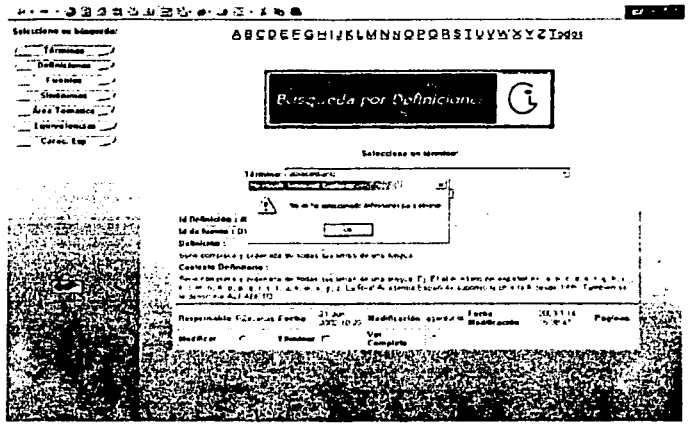


Figura 4-8. Mensaje de alerta por no seleccionar una definición

Por otro lado, si el usuario está seguro de eliminar la definición, el sistema nos advierte si en verdad queremos suprimir la definición, dándonos la opción de cancelar sin dañar la información. En caso de afirmarle que se desea eliminar la definición, automáticamente se borrará del banco de forma definitiva y ya no habrá opción para recuperarla (Figura 4-9)

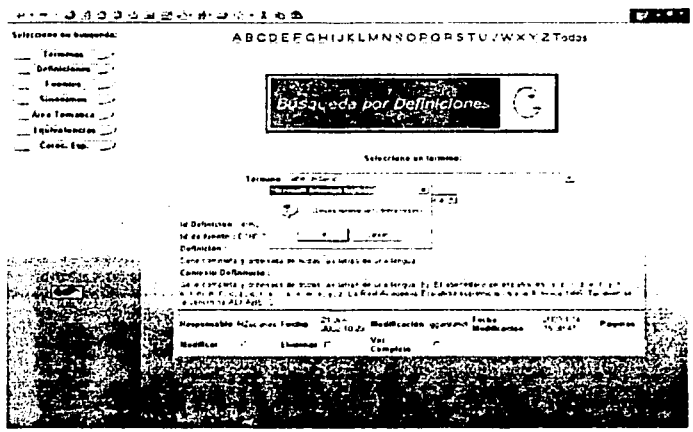


Figura 4-9. Confirmando la eliminación de la definición

En caso de que se requiera introducir caracteres especiales (<, >, ~, etc.), o algún símbolo del alfabeto griego (α, β, χ, etc.); se deberán seguir ciertas reglas que aparecen cuando damos clic en "Carac. Esp", la cual es una opción de ayuda para los usuarios.

El sistema nos presenta un manual de cómo se tiene que introducir los caracteres especiales (Figura 4-10). Los pasos son similares a los que seguimos para introducir la tipografía (negrillas, itálicas, etc.); salvo que en este caso introduciremos el código HTML indicado en la interfaz de "lista de caracteres especiales". Veamos el siguiente ejemplo: Si el usuario desea que aparezca el carácter especial "<", se deberá introducir la etiqueta "<".

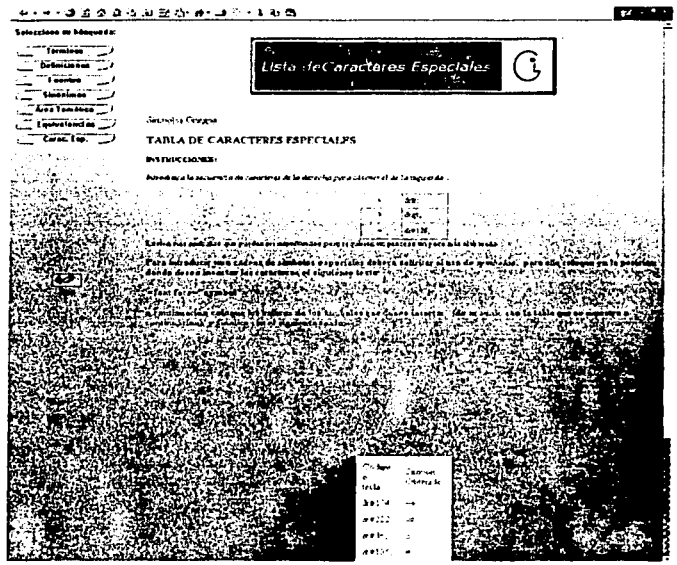


Figura 4-10. Caracteres especiales

Dentro de la interfaz de caracteres especiales se encuentra un enlace a la "tabla de símbolos del alfabeto griego" (Figura 4-11), que para este caso el proceso de etiquetado es diferente. Si se desea que aparezcan los símbolos "α, β y χ", y el sistema muestre tales símbolos, es necesario hacer uso de las siguientes etiquetas: A, B y G . Nótese que se hace uso de letras en mayúscula para hacerle la tarea más fácil al usuario.

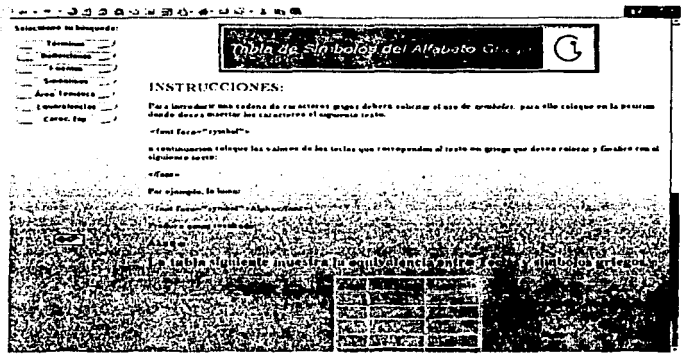


Figura 4-11. Alfabeto griego

4.2.3 Interfaz del usuario del banco

Con el rol de "usuario del banco", el sistema permite agregar, consultar y/o modificar la información, pero no eliminarla (Figura 4-12). A esta clase de usuario no se le presenta la información de los responsables de la inserción y modificación de la información, y de igual forma no se muestran las fechas debido a que son parte de la administración del banco.

Cuando el usuario desea eliminar ciertos datos del banco terminológico, el sistema automáticamente envía un correo electrónico con la petición de eliminación a los administradores del banco y del sistema (Figura 4-13), la cual será sujeta a aprobación por los expertos en el área.

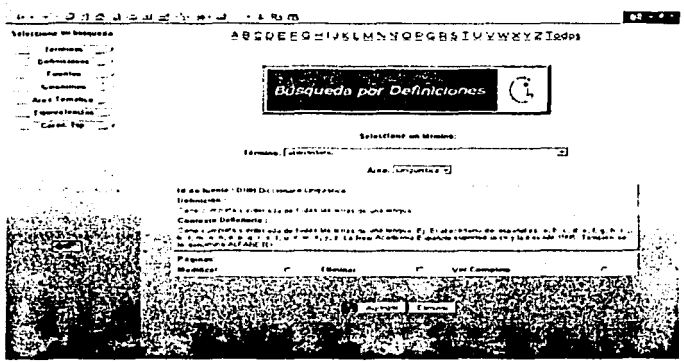


Figura 4-12. Consulta de una definición para el rol de "usuario del banco"

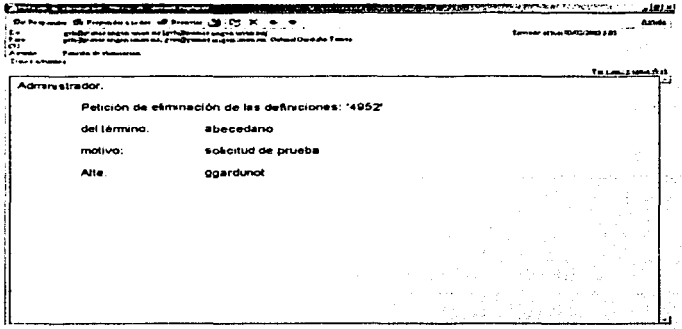


Figura 4-13. Correo con la petición de eliminación

4.2.4 Interfaz del visitante

Cuando el usuario accede al banco terminológico con el rol de "visitante", el sistema le muestra únicamente la información necesaria para que se documente. Este tipo de rol no tiene la posibilidad de agregar, modificar y/o eliminar la información del sistema, solamente estará disponible como consulta (Figura 4-14).

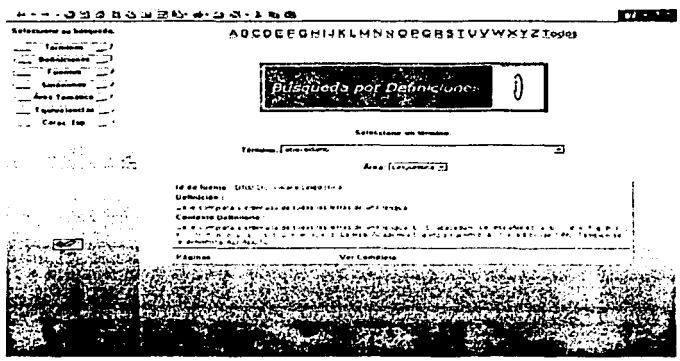


Figura 4-14. Consulta de una definición para el rol de "visitante"

4.2.5 Pruebas de funcionalidad

Para asegurar que el banco terminológico funciona perfectamente, es esencial invertir una cierta cantidad de tiempo elaborando las pruebas de funcionalidad. Hay que probar que el sistema sea útil, no sólo en un caso concreto, sino en todos. Podemos detectar fallos en dos niveles, principalmente:

- Errores en los datos de entrada, como por ejemplo, un valor fuera de rango.

- Errores en la lógica del programa, por ejemplo, prevenir las divisiones por cero.

A la hora de seleccionar los datos de prueba que se introducirán como entrada al programa para detectar posibles errores, haremos uso de estos dos métodos:

- Método de los valores extremos: en primer lugar buscamos comprobar que el sistema reacciona adecuadamente ante datos de entrada válidos que sabemos de antemano conducen a una solución conocida. A continuación, probamos con datos no válidos para comprobar la capacidad de detección de errores del programa y por último lo hacemos con datos reales.

- Método perfilador: se eligen datos de prueba para que cada línea de código se ejecute al menos una vez. Si existe alguna parte de código que, de acuerdo con una determinada condición, deba saltarse totalmente, nos aseguramos que existe al menos una entrada de prueba para que sí la ejecute.

Por otro lado se le hicieron pruebas de seguridad al sistema de tal forma que se tuviera la certeza de que no accederá al banco terminológico algún usuario que no tenga autorización de hacerlo. Así mismo, se verificó que los usuarios que pueden hacer uso del banco, no puedan manipular la información (agregar, modificar y/o borrar), si es que no cuentan con los permisos necesarios.

4.2.6 Mantenimiento

Una vez que el banco terminológico estuvo concluido, se asignaron permisos de acceso a los usuarios para que éstos comiencen a darle uso y posiblemente surjan dos causas fundamentales por las cuales el sistema deba entrar en fase de mantenimiento:

- Producción de errores que no se descubrieron durante la fase de prueba: Aún si la fase de prueba ha sido concienzudamente explotada, es posible que no hayamos tomado en cuenta alguna situación en la que el sistema no funcione correctamente.

- Corregir, mejorar o adecuar el sistema debido a la aparición de nuevo software y/o hardware, o incluso por cambios en las necesidades de los usuarios, en cuyo caso será imprescindible volver a las fases de análisis y diseño.

Capitulo 5. Aplicaciones del banco terminológico como base de conocimiento

Dentro del campo de la inteligencia artificial, la construcción de sistemas expertos inteligentes se basa en la posibilidad de que las computadoras posean una cierta capacidad de razonamiento basada en el conocimiento y la experiencia.

Los especialistas en teoría del conocimiento, psicología y lingüística han trabajado conjuntamente para describir la forma en que los seres humanos adquieren, estructuran y procesan el conocimiento, además de cómo se retroalimentan con la experiencia, con el fin de construir un modelo que represente la actuación humana[1].

Los expertos en cómputo e inteligencia artificial se han propuesto implantar en una computadora el modelo elaborado por los expertos, de manera que la máquina actúe de forma inteligente, que sea capaz de reproducir e interpretear lo que se observa en la realidad, que comunique pensamientos, razone y tome desiciones.

El conjunto de términos de un área especializada se comporta como una estructura de conceptos, la cual refleja la organización del conocimiento sobre el área en cuestión. De esta manera, los términos se convierten en piezas clave de la representación del conocimiento especializado. Un sistema experto inteligente necesita tales estructuras conceptuales para procesar la información, y actuar de acuerdo con lo que se le solicita; por esta razón irá adquiriendo su base de conocimiento a través de los términos.

Como los sistemas expertos no trabajan únicamente con un área especializada, sino con varias áreas a la vez, necesitan la terminología empleada dentro de éstas. La forma de adquirir y mantener tal información es a través de los *bancos terminológicos*.

Ahora bien, el banco terminológico objeto de esta tesis, constituye una base de conocimiento particular para el Grupo de Ingeniería Lingüística (GIL), ya que sus actividades dependen en gran medida de la información contenida en el banco terminológico. Especialmente, son tres las aplicaciones que cabe desatacar: contextos definitorios, palabras clave y paradigmas semánticos, los cuales son descritos a continuación.

5.1 Contextos definitorios

En este momento, el GIL se encuentra desarrollando una línea de investigación que consiste en elaborar una herramienta que extraiga de forma automática los contextos definitorios contenidos en los textos especializados, tales como revistas científicas, informes técnicos, tesis, etc.[2] Un *contexto definitorio* es un fragmento textual donde se

[1] Cabré (1993)
[2] Alarcón (2003)

introduce un término, su correspondiente definición y algún elemento característico: marcas tipográficas, predicaciones verbales y predicaciones pragmáticas.

Tal investigación tiene como finalidad, entre otras cosas, elaborar diccionarios o glosarios correspondientes a un área determinada. Para llevarlos a cabo, el terminólogo debe identificar, en una primera etapa de su trabajo, aquellos contextos definitorios que presenten un posible término junto con su definición. Esta etapa del trabajo terminológico consume un tiempo considerable y actualmente se desarrolla de forma manual.

Si bien la ingeniería lingüística ha desarrollado instrumentos que permiten extraer automáticamente los posibles términos representativos de un documento[3], aún es necesaria la elaboración de una herramienta capaz de reconocer los posibles conceptos de un texto especializado, es decir, los términos y sus correspondientes definiciones. El alcance de un sistema de extracción automática de contextos definitorios no se limita al trabajo terminológico, sino también es útil en otras áreas de estudio, tales como la enseñanza de lenguaje especializado y la inteligencia artificial.

Se ha observado que cuando un autor introduce un término que no es del todo conocido por los lectores, o bien define y detalla una serie de información adyacente al término, o utiliza una serie de patrones tipográficos y sintácticos que le permite resaltar dicha información. En el siguiente ejemplo se observa cómo el autor define el término "hospital" con ciertas marcas tipográficas (comillas, cursivas y negritas) y al mismo tiempo muestra la referencia bibliográfica.

> Según G. Malagón (1996, p.18) un hospital se define como: *"una parte integrante de la organización médica, cuya función es la de proporcionar a la población atención médica completa, tanto preventiva como curativa y cuyo servicio de consultorio externo alcanza a la familia en el hogar..."*

5.1.1 Patrones recurrentes en contextos definitorios

Durante el proceso de su investigación, Alarcón[4] encontró una serie de patrones recurrentes en los contextos definitorios, los cuales agrupó en cuatro grupos distintos, que van de formas simples a complejas: tipográficos, sintácticos, mixtos y compuestos. Utilizó cierta simbología para facilitar el reconocimiento de los patrones y sistematizar el análisis. Los símbolos utilizados son: <T>→ Término, <D> → Definición, <MT> → Marca tipográfica, <P1> → Predicación pragmática y <P2> → Predicación verbal.

- **Patrones Tipográficos:** Los patrones tipográficos son aquellos en donde sólo intervienen factores de formato de texto para enfatizar ya sea el término, la definición o ambos. Entre las formas tipográficas características se encuentran palabras en mayúscula, cursivas, negritas, subrayadas, o bien introducidas por una viñeta o con signos de puntuación (Tabla 5-1).

[3] Cabré (2001)
[4] Alarcón (2003)

Tabla 5-1. Ejemplo de un patrón tipográfico

Contexto Definitorio
<T><MT>*IMPACTOS AGREGADOS BIOLOGICOS*<MT> <MT>.<MT> <D>Los que resultan al sistema biológico...<D>

- **Patrones Sintácticos:** Se consideran como patrones sintácticos aquellos que no presentan ningún tipo de relación tipográfica en su forma. En estos patrones, los elementos característicos son las predicaciones pragmáticas o verbales (Tabla 5-2).

  En las *predicaciones pragmáticas* se incluye aquella información adyacente al término, como lo es el autor, el contexto de alcance del término propuesto, el área a la cual pertenece el término, etc.

  Las *predicaciones verbales* son aquellas palabras que conectan al término con la definición. Una forma muy usual que emplean los autores es el pronombre "se" más un verbo como definir, entender, considerar, referir, etc.

Tabla 5-2. Ejemplo de un patrón sintáctico

Contexto definitorio
<T>Un Soporte Logístico de Plataforma<T> <P1>de manera general<P1> <P2>se define como<P2> <D>un territorio equipado para el desarrollo de actividades logísticas...<D>

- **Patrones Mixtos:** En los patrones mixtos se presentan las dos características anteriores, esto es, están formados por una secuencia donde el término, la definición, o bien ambos son resaltados mediante alguna marca tipográfica y además incluyen alguna predicación verbal o pragmática (Tabla 5-3).

Tabla 5-3. Ejemplo de un patrón mixto

Contexto definitorio
<P1>Según G. Malagón (1996, p.18)<P1> <T>un hospital<T> <P2>se define como<P2>: <D><MT>*"una parte integrante de la organización médica, cuya función es la de proporcionar a la población..."*<MT><D>

- **Patrones Compuestos:** Los patrones compuestos pueden ser de dos tipos: 1) un mismo contexto definitorio sirve para introducir dos o más términos distintos y 2) la definición de un término sirve, a su vez, como contexto definitorio para la introducción de un nuevo concepto (Tabla 5-4).

Tabla 5-4. Ejemplos de patrones compuestos

Contexto definitorio
<P1>Se considera<P1> <T1>calamidad<T1> <D1>todo acontecimiento que pueda impactar el sistema afectable<D1>, <P1>en este caso<P1> <T2>la central y sus alrededores<T2>, <D2>incluyendo la mina Carbón II...<D2>

5.1.2 Aspectos computacionales

El banco terminológico con el que cuenta el GIL tiene un campo que está destinado a los contextos definitorios de donde provienen los términos (4.2.2). Este campo resulta de suma importancia para, por un lado, obtener las definiciones que están contenidas en el mismo banco, como, por el otro lado, para el desarrollo del programa de extracción automática de términos y definiciones de textos de especialidad. En primera instancia, el banco terminológico es un repositorio de los contextos definitorios encontrado en los documentos y, por tanto, este campo está diseñado de forma que se puedan conservan las marcas tipográficas por medio de etiquetas HTML, con lo cual nos estamos asegurando de no perder el formato original que le dio el autor al documento.

Posteriormente, el programa de extracción automática de conceptos, que se está analizando y diseñando en la ya mencionda línea de trabajo, utilizará el banco terminológico como proveedor principal de información. Los contextos definitorios serán extraídos del banco por medio de un programa lingüístico, que a su vez los analizará para detectar los diferentes patrones recurrentes que pueda contener.

El motivo de hacer un análisis exhaustivo para encontrar los diferentes patrones en los contextos, nos facilita la localización y extracción de los términos, de sus respectivas definiciones y de los elementos característicos (marcas tipográficas, predicaciones verbales y predicaciones pragmáticas). Todos estos datos se almacenarán en una nueva base (Figura 5-1), actualmente en desarrollo.

En esta etapa se guardarán los términos, las definiciones, las predicaciones verbales y las pragmáticas, así como las marcas tipográficas por separado, ya que estarán sujetas a pruebas con el fin de analizar qué tan fiable es esta parte de la herramienta de extracción automática. Los términos y definiciones con las que se compararán, podrán ser los que contiene el propio banco terminológico. La información contenida en el banco es de completa fiabilidad, debido a que fue analizada y catalogada por expertos en el área.

Una vez que se hayan elaborado las pruebas pertinentes y se tenga la certeza de que el programa lingüístico funciona correctamente, se harán las modificaciones necesarias para que ya no exista la necesidad de extraer los contextos del banco terminológico; sino pasar directamente al texto original completo previamente digitalizado, con el fin de detectar y analizar los posibles contextos definitorios, para luego aplicar el programa que detecte los patrones existentes. De este modo, se obtendrán automáticamente los términos y sus definiciones. Al final, se almacenará automáticamente la información dentro del banco terminológico (Figura 5-2).

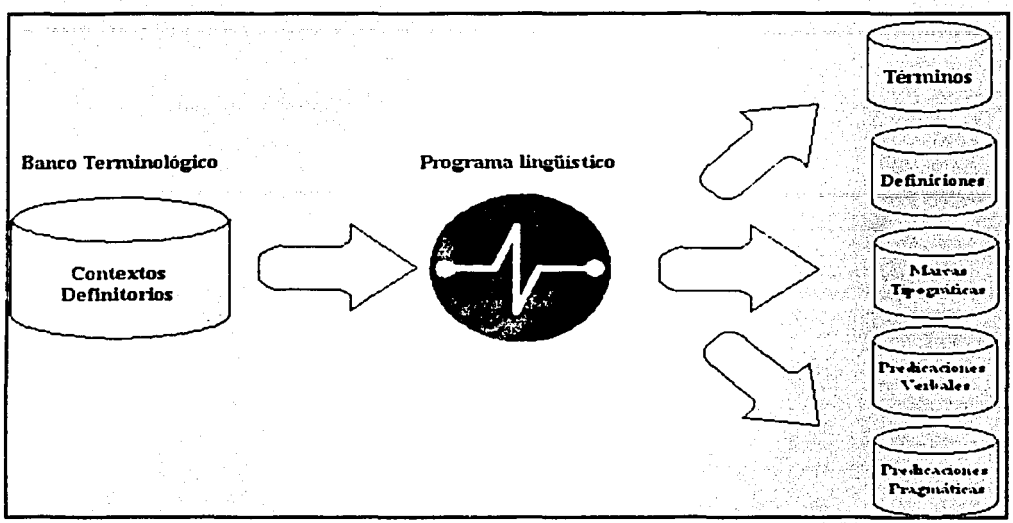


Figura 5-1. Arquitectura de prueba

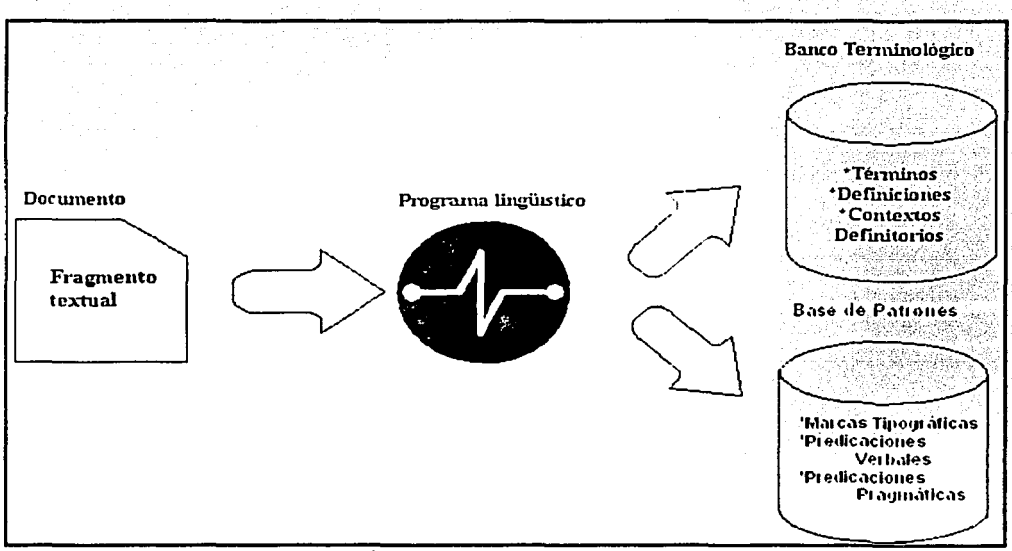


Figura 5-2. Arquitectura del extractor automático de contextos definitorios

5.2 Palabras clave

El principal proyecto de investigación del GIL es la elaboración de un diccionario onomasiológico (2.1), del cual se desprenden diferentes líneas de investigación. Una de ellas consiste en la elaboración de la *Base de Conocimiento Léxico* (Figura 5-3[5]), la cual es parte integral del diccionario. La base de conocimiento consta de dos bloques: el primer bloque concierne al banco terminológico, y el segundo bloque lo constituye los términos, las palabras clave y los paradigmas semánticos.

El término *palabra clave* lo usamos para propósitos de recuperación de información y con éste se designa cualquier palabra relevante usada en la consulta que genera el usuario. Cualquier palabra contenida en las definiciones puede ser significativa y nos puede guiar dentro del concepto escrito por el usuario; de esta manera, se recupera el término asociado al concepto introducido.

Contrario a las palabras clave, las *palabras funcionales* no son relevantes en la recuperación de información, pero son muy importantes para conectar unas palabras clave con otras y hacer el concepto comprensible. En recuperación automática, las palabras funcionales comunes y cualquier palabra irrelevante puede ser eliminada vía un *stop list*, el cual contiene una lista de las palabras que serán ignoradas en el procesamiento de la información, tales como artículos, pronombres, algunos adjetivos, etc.

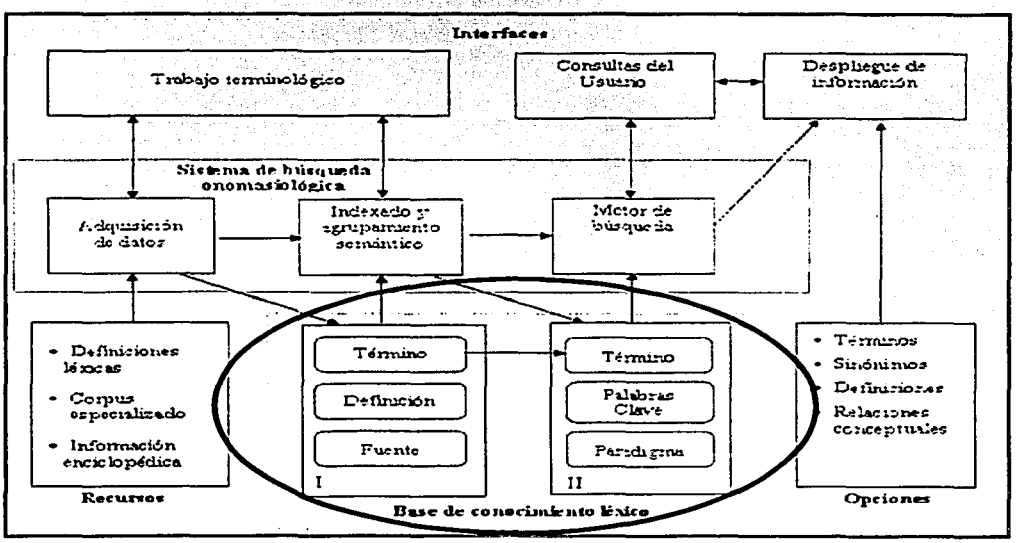


Figura 5-3. Arquitectura del diccionario onomasiológico

[5] Sierra y McNaught (2000a)

El objetivo principal de esta línea de investigación, es generar las palabras clave de forma automática con ayuda de la herramienta que se elaboró en el GIL. El fin radica en asociar las palabras clave con sus paradigmas semánticos que le corresponden.

5.2.1 Herramienta generadora de palabras clave

El éxito de una búsqueda onomasiológica depende de la precisión en la detección de todas las palabras clave contenidas en el concepto. Dado que el usuario a menudo no emplea las palabras precisas que concuerden con las palabras clave, la recuperación de términos puede ser lejana del concepto introducido; por ello, el GIL desarrolló una herramienta que ayuda al experto a encontrar las palabras clave de forma automática que se encuentran contenidas en las definiciones.

El funcionamiento del programa consiste en hacer una conexión al banco terminológico y extraer todas las definiciones pertenecientes al término en cuestión. Al mismo tiempo, se consulta un archivo en donde se encuentra el *stop list*. La elaboración del *stop list* fue analizada y descrita minuciosamente, con el fin de evitar la eliminación de palabras clave al considerarlas como palabras funcionales.

Una vez que se cuenta con las definiciones y el *stop list*, se utiliza un módulo del programa que genera las palabras clave propuestas por el software. Un experto en el área se da a la tarea de hacer una revisión de las palabras clave dadas por el sistema y acepta las que serán válidas. Una vez validadas son ingresadas en una base de datos (Figura 5-4).

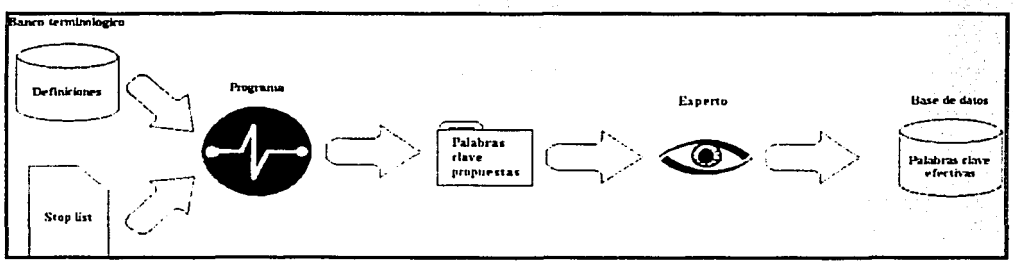


Figura 5-4. Arquitectura generador de palabras clave

5.2.2 Interfaces del sistema

A continuación se hablará a detalle en que consiste la aplicación "generador de palabras clave", sus interfaces y funcionamiento, como apoyo en la elaboración del diccionario onomasiológico.

En primera instancia, el sistema nos da la opción de elegir el término al cual se requiere generarle sus palabras clave a partir de sus definiciones contenidas en el banco terminológico. También nos muestra un enlace a "Ver todos los archivos generados" que

se explicará mas adelante (Figura 5-5). En las figuras que se presentan en esta sección se muestran ejemplos aplicados al área de física.

Una vez que se eligió el término, el sistema hace la conexión al banco terminológico y extrae todas las definiciones que le pertenecen al término en cuestión, toma el archivo con el *stop list* y elimina todas las palabras funcionales. La figura 5-6 nos muestra el término que se escogió, que para efectos de mostrar los resultados del sistema, utilizaremos el término "acción". La interfaz nos presenta todas las palabras clave propuestas que generó el sistema, con la capacidad de poder visualizar cada una de ellas en todas las definiciones (opción a la derecha). Por otro lado, en primera instancia nos proporciona todas las palabras marcadas como efectivas (cajas verificadas a la izquierda), ya que el sistema propone que todas las palabras aquí presentadas son efectivas, a menos que el experto en la materia opine lo contrario.

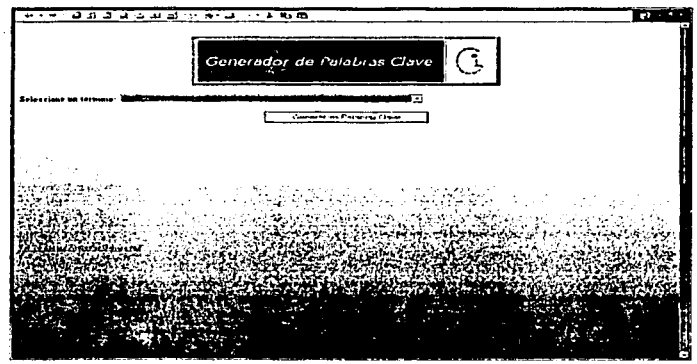


Figura 5-5. Interfaz principal

Si el experto está seguro que todas son palabras clave efectivas, puede guardarlas en este momento; si no es así, tiene la opción de pasar a dos diferentes opciones: 1) ver las palabras clave seleccionadas en todas las definiciones, y 2) ver las palabras clave seleccionadas en todos los contextos definitorios. Si lo que quiere es salir del programa, puede hacerlo pulsando en el botón de "Menú principal", y con ello regresa a la interfaz principal sin haber guardado nada.

Cuando el experto está en el proceso de verificación de las palabras clave, puede suceder que alguna de ellas cause incertidumbre sobre el hecho de que lo sea o no, por lo que al hacer clic en la opción del lado derecho se puede visualizar la palabra contenida en todas las definiciones. Esta opción nos presenta todas las otras palabras que rodean a la palabra clave para ver su contexto en la definición, y de esta manera clarificar lo confuso de la palabra (Figura 5-7). Se puede regresar a la interfaz anterior para continuar con el proceso de revisión o simplemente ir a la interfaz principal.

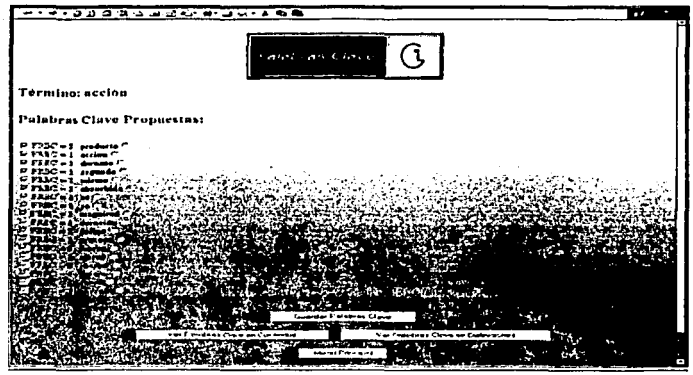


Figura 5-6. Palabras clave propuestas

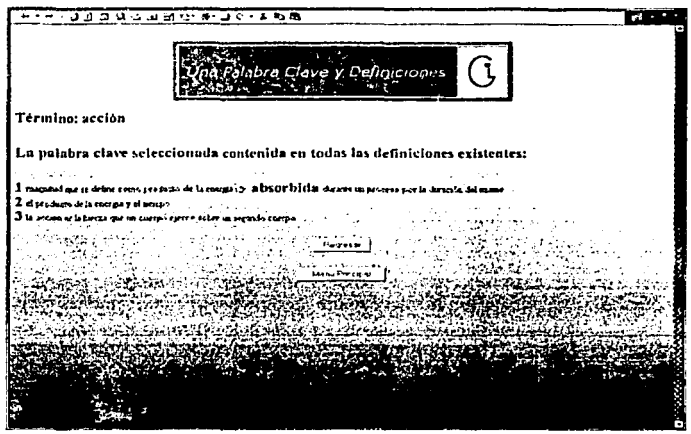


Figura 5-7. Una palabra clave en todas las definiciones

Como ya se hizo mención, el éxito de encontrar el término buscado por medio del diccionario onomasiológico, depende de encontrar y clasificar correctamente las palabras clave. Cuando el experto termina de seleccionar las palabras clave, puede verificarlas, ya sea en todas las definiciones (Figura 5-8), o a todas ellas inmersas en los contextos definitorios que le pertenecen (Figura 5-9). De esta forma, se puede dar cuenta si en verdad todas son palabras clave efectivas o si es necesario marcar o descartar alguna. A partir de estas dos interfaces se puede guardar la información sin necesidad de regresar a la pantalla anterior.

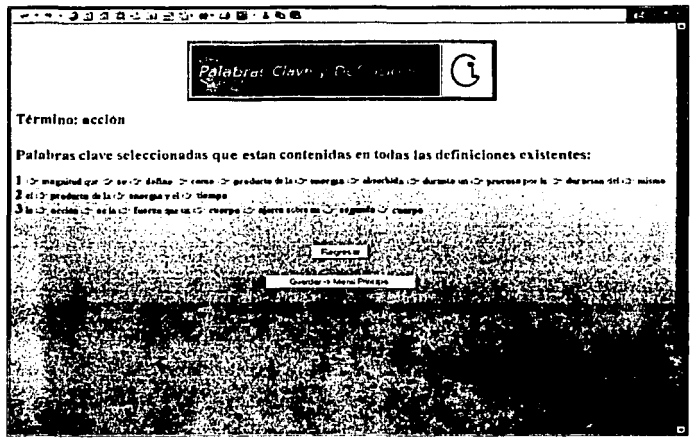


Figura 5-8. Palabras clave seleccionadas en todas las definiciones

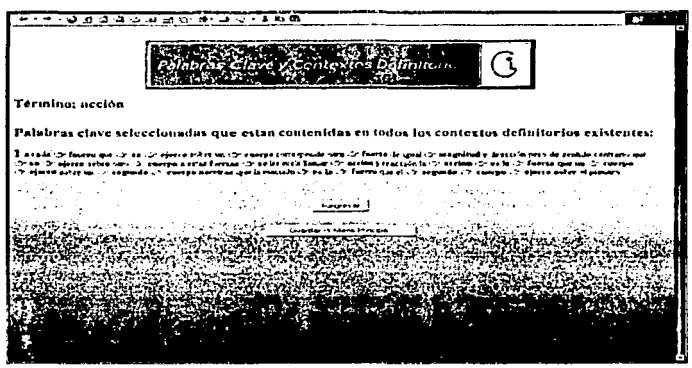


Figura 5-9. Palabras clave seleccionadas en todos los contextos definitorios

En la interfaz principal (Figura 5-5), el usuario puede acceder a los archivos generados por medio del enlace que se nos presenta. Cuando pulsamos en dicho enlace observamos una pantalla como la mostrada en la figura 5-10, que nos indica la fecha en que fue creado el archivo, el tamaño del archivo y el nombre. Se puede observar que todos los nombres son diferentes dado que se decidió nombrarlos conforme se denomina el término analizado: esto facilita la búsqueda de la información. Si damos clic en el nombre del archivo, nos muestra todas las palabras clave que están relacionadas al término en cuestión.

Finalmente todas las palabras clave efectivas seleccionadas y verificadas, se guardan en una base de datos, ya que, como hemos mencionado, forman parte de la base de

conocimiento léxico, además de ser el punto de partida de otra línea de investigación enfocada a la localización de los paradigmas semánticos asociados a las palabras clave. La herramienta de apoyo a la determinación de palabras clave puede ser consultada en http://imsai.iingen.unam.mx/diccionarios/perl/kw1.pl

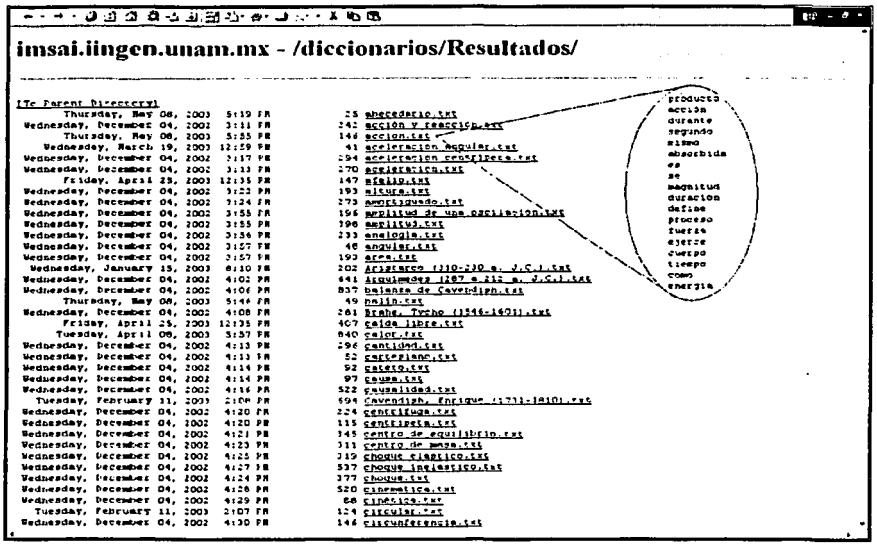


Figura 5-10. Archivos generados

5.3 Paradigmas semánticos

En el contexto del diccionario onomasiológico, formalizar un concepto con las palabras clave exactas es algunas veces una dura tarea para el usuario, pero las búsquedas pueden llegar a ser más arduas si el usuario también tiene que identificar los paradigmas relacionados a las palabras clave, particularmente cuando la consulta es hecha en lenguaje natural.

Por ello, una línea de investigación dentro del GIL consiste en extraer automáticamente los paradigmas semánticos a partir de pares de definiciones[6]. Esta aplicación sigue formando parte de la *Base de Conocimiento Léxico*, que a su vez es parte integral del diccionario onomasiológico (Figura 5-3).

Este trabajo necesita hacer uso del banco terminológico que se elaboró en la presente tesis, con el fin de transferir directamente la información necesaria, y encontrar así los paradigmas semánticos. Por otro lado, es necesario contar con las palabras clave que les

[6] Sierra y McNaught (2000b)

pertenecen a los términos, las cuales se obtienen a partir de la herramienta mencionada en el apartado anterior (5.2).

Durante la búsqueda de la definición perteneciente a un término, en el diccionario onomasiológico, el concepto proporcionado por el usuario nos puede conducir a un resultado inadecuado. Veamos la siguiente definición:

> **Telescopio**: 1. m. *Ópt.* Instrumento que permite ver agrandada una imagen de un objeto lejano. El objetivo puede ser o un sistema de refracción, en cuyo caso el telescopio recibe el nombre de anteojo, o un espejo cóncavo[7].

Si el usuario escribe: *"Instrumento para ver objetos lejanos"* el sistema identificará **Telescopio**, pero si escribe *"Aparato para observar objetos lejanos"* el sistema no sabrá que término proporcionar.

Ahora bien, la forma de resolver este problema es expandir la búsqueda de palabras clave, es decir encontrar los paradigmas semánticos que le pertenecen a cada una de ellas. Los *paradigmas semánticos* son palabras semánticamente relacionadas, esto es, que al cambiar una palabra por otra, *Instrumento* → *Aparato* o *Aparato* → *Instrumento*, el sentido de la definición sigue siendo el mismo

Si se generan los siguientes paradigmas semánticos: *aparato, utensilio, artefacto, instrumento, útil* como parte del mismo conjunto de palabras clave interrelacionadas, la descripción proporcionada por el usuario se expandirá a tantas opciones como paradigmas encontrados existan:

- ***Aparato*** para observar objetos lejanos.
- ***Utensilio*** para observar objetos lejanos.
- ***Artefacto*** para observar objetos lejanos.
- ***Instrumento*** para observar objetos lejanos.
- ***Útil*** para observar objetos lejanos.

El método empleado identifica los paradigmas semánticos de forma automática a partir del alineamiento de dos definiciones que forzosamente tienen que ser de fuentes distintas, ya que si no es así, no se estarían tomando fuentes disímiles, sino diferentes acepciones de la misma fuente. El alineamiento de cadenas es aplicado con gran aceptación en distintas áreas tales como la inteligencia artificial, el reconocimiento de voz, la recuperación de información, la traducción automática, la corrección de errores en la transmisión de mensajes a través de un medio, etc.

[7] DRAE electrónico

5.3.1 Herramienta generadora de paradigmas semánticos

Hay que tomar en cuenta que en la búsqueda onomasiológica, al igual que en la mayoría de los sistemas de recuperación de información, la efectividad de la búsqueda se incrementa considerando los lemas de las palabras introducidas por el usuario. El sistema elaborado para determinar los paradigmas semánticos cuenta con un lematizador que permite la identificación de los lemas en forma automática. La función de un lematizador es detectar todas las palabras que tienen la misma raíz y asignarles el mismo lema a todas, por ejemplo:

Palabras: cálculo, calcular, calculadora, calculará, calculando, calcular...
Lema: calcul

Ahora bien, la herramienta de determinación de paradigmas semánticos obtiene, por cada término del banco terminológico, un par de definiciones que provengan de distintas fuentes y que le correspondan (Figura 5-11)[8]. Lematiza las definiciones, las alinea, las compara e identifica los pares vinculados, es decir, pares de paradigmas semánticos. Si el término cuenta con más de dos definiciones, irá comparando cada una con todas las demás, generará todos los pares vinculados de todas las combinaciones de definiciones, agrupará todos los pares de paradigmas y finalmente ingresará en una base de datos todos los paradigmas generados. El sistema termina su ejecución cuando ya no encuentre más definiciones del mismo término.

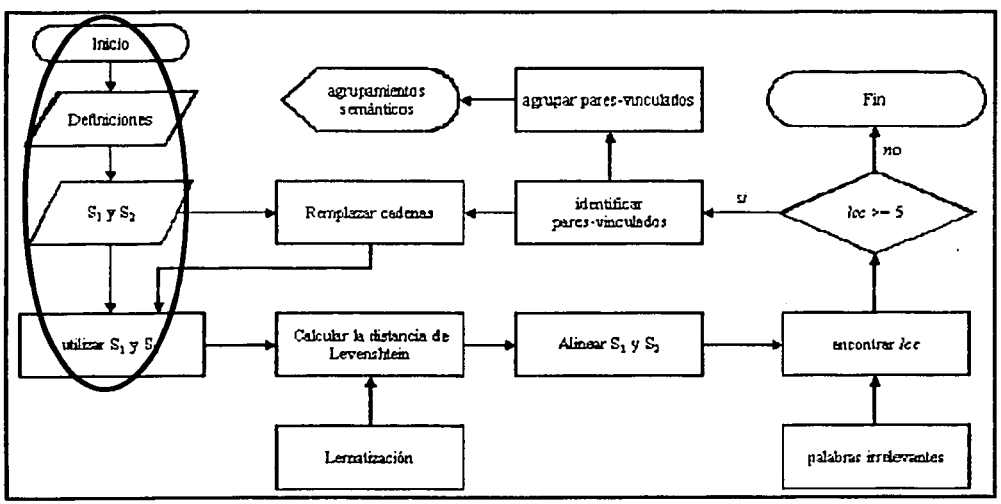


Figura 5-11. Algoritmo de paradigmas semánticos

[8] Castillo (2002)

5.3.2 Interfaces del sistema

Enseguida se explicará el funcionamiento de la herramienta "Clustering System: Alineamiento Semántico" como parte integral de la elaboración del diccionario onomasiológico y su interrelación con el banco terminológico.

En principio, el sistema hace conexión con el banco terminológico y nos brinda la opción de seleccionar la base terminológica de la cual se desean extraer los paradigmas semánticos. La figura 5-12 nos muestra la interfaz de entrada al programa que señala las diferentes bases en las que se puede trabajar, siendo éstas: lingüística, física, fenómenos destructivos, metrología, desastres, ingeniería, arabismos, ingeniería lingüística y sexualidad. En principio aparece primero la de lingüística. Por otro lado, se puede elegir la cantidad de definiciones que se desean alinear: por default nos coloca las primeras 20 definiciones, pero se puede trabajar con las *n* definiciones que contenga el banco terminológico. Dado que algunas bases se encuentran en inglés y otras en español es necesario contar con un lematizador para cada idioma, por lo que el sistema nos permite lematizar o no las definiciones y, en caso afirmativo, escoger el idioma del lematizador.

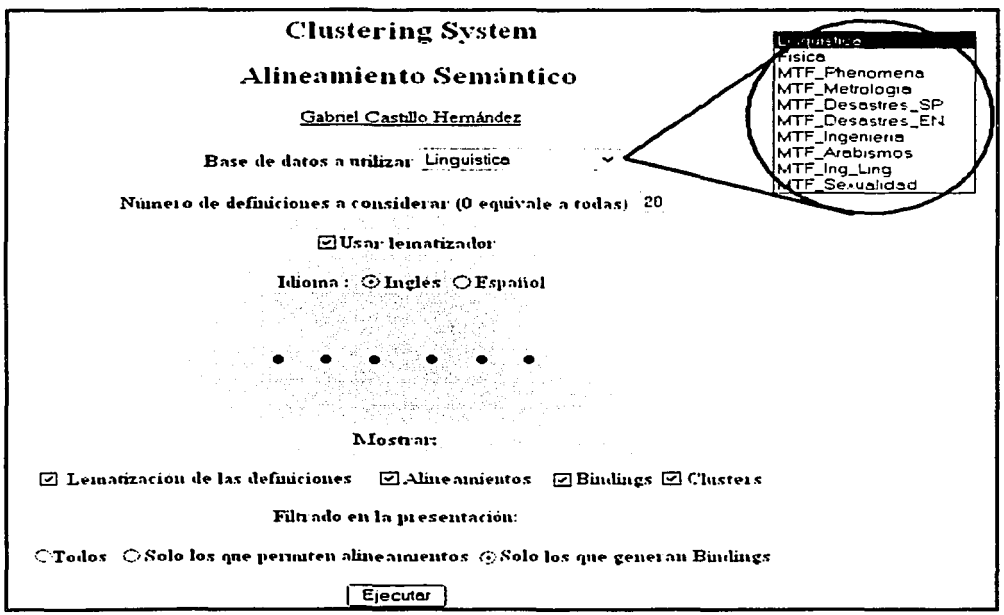


Figura 5-12. Interfaz principal de la herramienta de extracción de paradigmas semánticos

En la interfaz aparecen otras opciones, las cuales son necesarias para el cálculo de los procesos en el sistema y así encontrar los paradigmas semánticos, cuya descripción no es relevante para esta tesis y puede ser consultada en Castillo (2002).

En la figura 5-12 se observa en la parte inferior los resultados que pueden obtenerse, en donde el sistema nos da las opciones de:

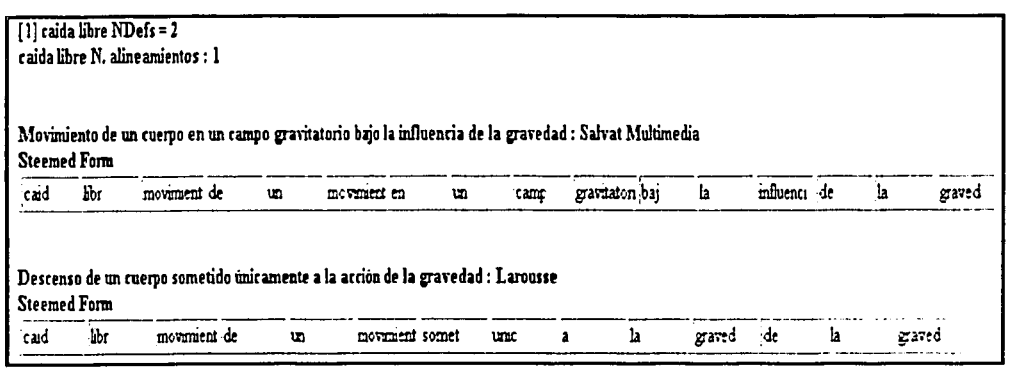
- Lematización de las definiciones.
- Alineamientos.
- Bindings (paradigmas semánticos).
- Clusters.

Entre estos resultados, el sistema puede filtrar la presentación con las siguientes opciones:

- Todas las iteraciones generadas.
- Sólo los que permiten alineamientos.
- Sólo los que generan bindings.

Ahora bien, una vez seleccionadas las distintas opciones, ejecutamos la herramienta, y el sistema hace la conexión a la base escogida y extrae las definiciones que le pertenecen al término en cuestión, junto con las fuentes de donde pertenece originalmente la información.

Para mostrar el funcionamiento de la aplicación, nosotros corremos el programa con la base de física y las cuatro opciones de salida. A continuación mostramos los resultados particulares para el término "caída libre". En la figura 5-13 se tienen las definiciones originales que se encontraron para el término caída libre, nos indica el nombre de la fuente a la que pertenece dicha definición y finalmente nos presenta la lematización para el término seguido de su definición.


[1] caida libre NDefs = 2
caida libre N. alineamientos : 1

Movimiento de un cuerpo en un campo gravitatorio bajo la influencia de la gravedad : Salvat Multimedia
Steemed Form

caid	libr	movment de	un	movment en	un	camp	gravitaton baj	la	influenci de	la	graved

Descenso de un cuerpo sometido únicamente a la acción de la gravedad : Larousse
Steemed Form

caid	libr	movment de	un	movment somet	unic	a	la	graved de	la	graved

Figura 5-13. Definiciones con sus lemas correspondientes

En la figura 5-14 se ilustra la forma en que se alinearon las definiciones para encontrar los posibles paradigmas semánticos (posibles bindings). Todas las palabras funcionales no podrán ser consideradas como paradigmas, pero no se descartan porque son vitales para encontrar los que realmente son paradigmas semánticos. La información adicional que es necesaria para el cálculo de los procesos en el sistema se explica en Castillo (2002).

Fuente Def 1 [Fte 0]	Fuente Def 2 [Fte 1]																
Def 1	caida	libre	movimiento	de	un	cuerpo	en	un	campo	gravitatorio	bajo	la	influencia	de	la	gravedad
Def 2	caida	libre	descenso	de	un	cuerpo	sometido	únicamente	a			la	acción	de	la	gravedad
Costos	0	0	1	1	1	1	2	3	4	5	6	6	7	7	7	7
Posible binding	NO	NO	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO	NO
lcc	0	0	6	0	0	0	4	1	1	0	0	0	5	0	0	0
Tipo	equal	equal	matched	equal	equal	equal	matched	matched	matched	null	null	equal	matched	equal	equal	equal

Figura 5-14. Alineamiento de las definiciones

Finalmente, nos presenta una tabla con los paradigmas semánticos resultantes por el par de definiciones alineadas anteriormente, junto con los clusters generados por el sistema (figura 5-15). Los *clusters* son el conjunto de palabras clave interrelacionadas mencionadas en 5.3. La herramienta puede ser consultada directamente vía Internet en la dirección:
http://tabasco.torreingenieria.unam.mx/scripts/clusters.exe/home

Bindings

Def 1	movimiento	influencia
Def 2	descenso	acción
Costos	1	7
Posible binding	SI	SI
lcc	6	5
Tipo	matched	matched

Clusters no lematizados

1 : movimiento descenso

2 : influencia acción

Figura 5-15. Bindings y clusters generados por el sistema

Conclusiones

En el presente trabajo se llevó a cabo una investigación a profundidad acerca de los bancos terminológicos, desde sus orígenes hasta los diferentes tipos de uso que pueden tener. En específico, el banco elaborado en esta tesis sirve de apoyo en las diferentes líneas de investigación que se llevan a cabo en el Grupo de Ingeniería Lingüística (GIL).

En el capítulo 1 se llegó a la conclusión que los *bancos terminológicos* son una colección de bases de datos que contienen las terminologías con sus respectivas definiciones sobre una temática especializada, que puede ser de carácter científica y/o tecnológica. Su misión es fijar y difundir el vocabulario especializado, además de ser utilizados como parte integral de sistemas más complejos.

Existe una gran variedad de personas que hacen uso de los bancos y encontramos que, por un lado, pueden llegar a ser desde usuarios muy especializados en alguna temática en especial, como un desarrollador de sistemas expertos inteligentes o un traductor de lengua; y por el otro, son usuarios comunes y corrientes de un nivel básico de estudios que consultan un diccionario de especialidad o una enciclopedia.

Los bancos son funcionales en diferentes aspectos, mencionando algunos de ellos tenemos que los podemos aplicar como:

- Un lugar para almacenar las bases de conocimiento de la inteligencia artificial y los sistemas expertos.
- Una herramienta útil para los traductores y un instrumento en el que existe una gran cantidad de información disponible para los especialistas.
- Diccionarios especializados disponibles en Internet y CD-ROM.
- Una herramienta que sirve de apoyo a otros sistemas computacionales.

En la historia de los bancos terminológicos hemos visto que han pasado por tres etapas diferentes. En la primera generación (70's) se utilizaron los bancos únicamente como un lugar para almacenar la información recopilada pero existía incompatibilidad entre los sistemas, la migración de información era casi imposible, se invertía mucho tiempo en las consultas y recuperación de la información, y por si fuera poco se necesitaba un especialista para manejar el sistema; a raíz de ello surgen los bancos de segunda generación. Con la aparición de los de segunda generación (80's) se superaron todos los problemas que tenían los de la generación anterior, se comienza con la investigación del desarrollo de interfaces en lenguaje natural y se pueden consultar los bancos terminológicos vía Internet y CD-ROM. Por último, en la generación inteligente (actualmente) los bancos se están desarrollando en el marco de la inteligencia artificial y los sistemas expertos, de tal forma que la computadora utiliza su información inicial

como elemento de partida para realizar por su propia cuenta otras operaciones más complejas y con mayor semejanza a las que haría un ser humano.

Hemos encontrado que se pueden elaborar bancos terminológicos clasificados de diferentes formas: informativos, equivalentes, plurilingües, enciclopédicos, visuales, etc. Su desarrollo puede ser una mezcla de diferentes clasificaciones, esto va a depender de las necesidades que tenga el solicitante del banco. A partir de este momento se tuvo una idea concreta de hacia dónde queríamos llegar; ahora, en las siguientes partes de este trabajo, mencionamos cómo fue que fuimos logrando paso a paso la elaboración del banco terminológico.

Con la ayuda de una serie de preguntas frecuentes (FAQ, Frequently Asked Questions) que propuso Ventura Miranda en su tesis de licenciatura, pudimos delimitar, en el capítulo 2, los requerimientos necesarios para el desarrollo del banco terminológico, como parte integral de un diccionario que realiza búsquedas onomasiológicas, el cual se está desarrollando en el GIL. Se pudo dilucidar que el banco también está destinado a fungir como una herramienta de apoyo a las diferentes líneas de investigación que se están llevando a cabo en el mismo Grupo.

El banco elaborado es de segunda generación, debido a diferentes aspectos que se consideraron:

- La terminología que contiene el banco, antes de introducirla en él, fue sujeta a aprobación. Esta labor dependió de los expertos en el área en cuestión y con esto garantizamos que la información que está contenida en el banco es veraz, confiable y de actualidad.

- La arquitectura del banco se diseñó de tal forma que se obtuvo una lógica y una estructura efectiva y flexible. Este trabajo ayudó a que el tiempo sea mínimo en la elaboración de una consulta y en la recuperación de información, además de que las consultas se realizan de forma autónoma vía Internet.

La clasificación del banco terminológico consta de seis puntos:

- Dado que la información es tal y como la presenta el autor en los textos de especialidad, el banco es descriptivo.

- El banco realiza las consultas semasiológicamente, es decir, se parte del término y el sistema nos guía hasta su definición.

- La temática encontrada en el banco es muy variada, por lo que es del tipo especializada general.

- Los contextos definitorios son extraídos de textos de carácter científico y/o tecnológico y de diccionarios especializados, lo que implica que los términos y su información adicional también proviene de ellos.

- El banco es monolingüe con equivalencias en otros idiomas, es decir, la información completa se encuentra en sólo un idioma (ingles o español), a excepción del término que tiene su equivalencia en otros idiomas.

- El banco terminológico está implantado en un servidor que cumple con los requerimientos mínimos para soportar las exigencias del banco.

La metodología que se utilizó para delimitar los requerimientos nos ayudó a concretar perfectamente el banco, conforme a las necesidades que se tenían, y ahora ya sabemos cómo se debe hacer el banco, qué funcionalidad va a tener y la ayuda que va a brindar al GIL al elaborarlo

El capítulo 3 se centra en dos partes fundamentales: el análisis y el diseño del banco terminológico. Con respecto al análisis encontramos que:

- La herramienta sirve de apoyo a las diferentes líneas de investigación que se están desarrollando en el GIL. La información recopilada por el GIL se encuentra clasificada y almacenada en un solo lugar y se puede consultar el sistema vía Internet.

- Los costos de software, asuntos legales, etc., los consideramos dados por hecho, ya que el sistema se elaboró dentro del GIL, en el seno del Instituto de Ingeniería de la UNAM, junto con el apoyo de CONACYT.

- Para la administración de los permisos, nuevos usuarios, nuevas bases terminológicas, permisos, etc., en el banco terminológico existe un DBA (Data Base Administrator) para llevar el control de los mismos. Por otro lado, el propio banco tiene tres usuarios con diferentes roles:

    a) El DBO (Data Base Owner) funge como administrador de una o varias bases a la vez y éste es el que decide los usuarios que pueden acceder a las bases y con qué rol. Este tipo de usuario puede consultar, agregar, modificar y/o eliminar información del banco.

    b) Se cuenta con un "usuario del banco" que está designado a alimentar las bases del banco y puede consultar, agregar y/o modificar la información del mismo, pero no eliminarla.

    c) Finalmente, existen los "visitantes" quienes tienen acceso únicamente a consultar el banco terminológico.

Por el lado del diseño del banco terminológico, se decidió utilizar, como sistema operativo, Windows 2000 Advanced Server®; como DBMS (Data Base Manager System) utilizamos a SQL Server 2000®; como servidor de Web, Internet Information Server (IIS)®; y los lenguajes de programación PERL, HTML y Javascript para la

elaboración de las interfaces del mismo. La decisión del sistema operativo y el DBMS se tomó porque son los recursos que se tenían disponibles en ese momento; el IIS viene incluido con el Windows y, con respecto a los lenguajes de programación, porque son compatibles con el sistema operativo y el DBMS y son de software libre.

Elaboramos un diagrama jerárquico funcional con el que pudimos clarificar todas las diferentes interfaces con las que cuenta el banco terminológico, y de esta manera conectar todos los caminos que interrelacionan cada una de las búsquedas.

Tomando como punto de partida el diseño que tenían las bases de datos originales, nos hicimos a la tarea de rediseñarlo con ayuda de la herramienta Power Designer® y el resultado que obtuvimos fue el modelo físico y conceptual de las bases de datos que conformaron el banco terminológico.

Gracias a esta parte de la investigación pudimos asegurar que se tuvieran controlados en las bases de datos: el almacenamiento de datos redundantes, la independencia entre los datos y las interfaces del banco.

En lo que concierne al capítulo 4, se procedió a la implantación del banco terminológico. En principio, se procedió a la elaboración de las bases de datos y a la programación de las interfaces del sistema, para después integrarlas, verificando que existiera una buena conexión entre las dos partes desde lugares remotos, es decir, desde máquinas distintas, conectadas al banco vía Internet.

Otro punto importante en este apartado fue la migración de la información. Dado que residía en bases de datos con cierta estructura, y el banco terminológico contiene otra, se tuvo la necesidad de apoyarnos en la hoja de cálculo de Microsoft Excel® para hacer los cambios pertinentes, siempre cuidando de no perder información.

Una vez implantado el banco y con información contenida en él procedimos a hacerle pruebas de funcionalidad para detectar las fallas en el sistema y corregirlas. Mencionando algunas de las pruebas que se hicieron: pruebas de detección de errores en la programación, pruebas de volumen, pruebas de conexión de varios usuarios a la vez, etc.

La fase de mantenimiento se ha puesto en marcha desde el término de las pruebas de funcionalidad. El banco se mantiene en producción hasta este momento y, dado que no es posible localizar todos los errores con las pruebas, los únicos que pueden encontrar las deficiencias del sistema son los usuarios; en ese aspecto tomamos en cuenta los comentarios de ellos para mantener el sistema siempre a la vanguardia.

Hasta esta fase de la investigación se han cumplido satisfactoriamente los dos primeros objetivos planteados en la presente tesis y el banco nos ha dado resultados que han llenado nuestras expectativas. Por un lado, se tiene un banco terminológico que es parte integral del diccionario onomasiológico y, por el otro, contamos con una herramienta que nos permite mantener y manipular toda la información terminológica recopilada por el GIL vía Internet.

Finalmente en el capítulo 5 se observó que los bancos terminológicos son de gran utilidad para muchas áreas de investigación, tal es el caso de la inteligencia artificial y los sistemas expertos que hacen uso de ellos para mantener actualizadas sus bases de conocimiento léxico. Con base en esto se pueden llegar a elaborar sistemas que se comporten como un ser humano.

Las aplicaciones que nos puede brindar el banco terminológico en las diferentes líneas de investigación que se están llevando a cabo en el GIL son inmensas. En este apartado hemos mencionado a tres de ellas (contextos definitorios, palabras clave y paradigmas semánticos) y fue vital el uso del banco para que se llegaran a concretar las diferentes líneas de trabajo. Con lo que respecta al tercer objetivo de esta tesis, se cumplió satisfactoriamente y la gama de utilidades que se le puede dar al banco terminológico es muy amplia, como ya lo hemos mencionado: de hecho, se está considerando al banco terminológico para que en un futuro forme parte de otras líneas de investigación.

Si bien el banco terminológico elaborado en la presente tesis fue desarrollado basándonos en uno de segunda generación, que es la más utilizada actualmente, falta investigar qué se ha llevado a cabo con los de generación inteligente. Uno de los retos es rediseñar el banco de tal forma que se comporte como un ser humano, ayudándonos de la inteligencia artificial y de los sistemas expertos.

Hay mucho por hacer en el desarrollo de bancos terminológicos; lo ideal es tener un banco que tenga la capacidad de contener imágenes, videos, sonidos, toda la información en diferentes idiomas, es decir, que contenga todas las clasificaciones existentes en los bancos y que el usuario tenga la oportunidad de adaptar el banco a sus propias necesidades, sin tener que estar rediseñando el banco cada vez que cambien los requerimientos. Además que el banco contenga interfaces en lenguaje natural y tenga la capacidad de razonar y aprender por sí mismo, tal y como lo hace un ser humano.

Bibliografía

[Alarcón, 2003] Alarcón Martínez, Rodrigo
"Análisis lingüístico de contextos definitorios en textos de especialidad"
Tesis de licenciatura,
Ciudad Universitaria, UNAM, 2003

[Cabré, 2001] Cabré, Teresa
"La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica"
Instituto Universitario de Lingüística Aplicada, Universidad Pompeu Fabra
Barcelona, 2001

[Cabré, 1999] Cabré, M. Teresa
"La terminología. Representación y Comunicación"
Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra
Barcelona, 1999

[Cabré, 1993] Cabré, M. Teresa
"La terminología. Teoría, metodología, aplicaciones"
Editorial Antártida/empúries
Barcelona, 1993

[Castillo, 2002] Castillo Hernández, Gabriel
"Algoritmo revisado para la extracción automática de agrupamientos semánticos"
Tesis de maestría,
Ciudad Universitaria, UNAM, 2002

[Cerdá, 1986] Cerdá Massó, Ramón
"Diccionario de lingüística"
Editorial Anaya
Madrid, 1986.

[Diccionario básico] "Diccionario básico de lingüística"
http://imsai.iingen.unam.mx/diccionarios/banco.htm

[DRAE electrónico] "Diccionario de la Lengua Española"
Edición electrónica, Versión 21.1.0
Real Academia Española, 1992

[Dore y Parina,1983] Dore, Dominique y Parina, Henridou
"Banco de datos. Utilización y funcionamiento"
Editorial Mitre
Barcelona, 1983

[Olguín.1997] Olguín, Heriberto
"Introducción a la cultura informática"
1ª edición, División de Ingeniería Eléctrica
Facultad de Ingeniería, UNAM. 1997

[Pressman, 1999] Pressman. Roger S.
"Ingeniería de Software"
3ª edición, Editorial McGraw Hill
1999

[Reyes,2002] Reyes Pérez, Antonio
"Hacia una obtención computarizada de términos. (Aplicación concreta al léxico de la física en el nivel bachillerato)"
Tesis de licenciatura,
Ciudad Universitaria, UNAM, 2002

[Sierra, 1999] Sierra Martínez, Gerardo
"Design of a concept-oriented tool for terminology"
Tesis de doctorado
University of Manchester
August, 1999

[Sierra y Alarcón, 2001] Sierra Martínez, Gerardo y Alarcón Martínez, Rodrigo
"Identificación de patrones recurrentes para la extracción automática de contextos definitorios"
Congreso Cuba 2001

[Sierra, Castillo, Reyes y Alarcón, 2001] Sierra, G., Castillo, G., Reyes, A. y Alarcón. R.
"Desarrollo de la ingeniería lingüística en la UNAM, México"
Congreso Jaén, España 2001

[Sierra y McNaught, 2000a] Sierra. Gerardo and McNaught, John
"Design of an onomasiological search system: A concept-oriented tool for terminology"
Terminology: International journal of theorical and applied issues in specialized communication.
John Benjamins Publishing Company, Amsterdam/Philadelphia
Volume6, number1, 2000

[Sierra y McNaught, 2000b] Sierra, Gerardo and McNaught, John
"Extracting semantic clusters from MRD's for an onomasiological search dictionary"
International journal of lexicography
Oxford University Press
Volume 13, number 4, December 2000

[Silberschatz, Korth y Sudarshan, 1998] Silberschatz, Abraham, Korth, Henry F. y Sudarshan, S.
"Fundamentos de Bases de Datos"
3ª edición, Editorial McGraw Hill
España, 1998

[Noyano y Fernández, 1998] Noyano Ávila, Antonio y Fernández Caballero, Antonio
"La documentación automatizada"
Editorial Librería Universidad
4ª Ed. Febrero 1998
Albacete, España

[Ventura,2002] Ventura Miranda, María Teresa
"La ingeniería de requerimientos como factor clave para el éxito de los proyectos de desarrollo del software"
Tesis de licenciatura,
FCA, UNAM
México DF, 2002

Páginas Web consultadas

http://atenea.udistrital.edu.co/profesores/jdimate/basedatos1/portada.htm

http://buscon.rae.es/diccionario/drae.htm

http://europa.eu.int/eurodicautom/login.jsp

http://muwa.trados.com

http://tis.consilium.eu.int/utfwebtis/frames/introfsEN.htm

http://www.euskadi.net/euskalterm

http://www.bibliodgsca.unam.mx/manuales/manual.pdf

http://www.cilf.org/bt.fr.html

http://www.colmex.mx

http://www.fmat.ull.es/~inform2/tema1.pdf

http://www.geocities.com/SiliconValley/Station/8266/perl/

http://www.granddictionnaire.com/_fs_global_01.htm

http://www.ilo.org/public/spanish/support/lib/dblist.htm

http://www.imf.org/external/np/term/index.asp

http://www.itu.int/search/wais/Termite/

http://www.lcc.uma.es/~av/Libro/CAP3.pdf

http://web.madritel.es/personales3/edcollado/ingsw/tema2.htm

http://www.microsoft.com/latam/sql/evaluation/overview/2000/datasheet.asp

http://www.microsoft.com/latam/sql/evaluation/overview/default.asp

http://www.microsoft.com/windows2000/es/advanced/help/

http://www.oup.com/elt/global/isbn/6890/

http://www.termium.com/site/espanol/index.html

http://www.upf.es/

http://www.uwasa.fi/comm/termino/

http://www.uwasa.fi/comm/termino/collect/index.html