Universidad Nacional Autónoma de México

Programa de Posgrado en Ciencias de la Administración

Tecnologías del lenguaje aplicadas a la administración de conocimiento

T e s i s

Que para optar por el grado de:

Maestro en Informática Administrativa

Presenta:
Juan Luis Serralde Galicia

Tutor:
Dr. Carlos Francisco Méndez Cruz
Facultad de Contaduría y Administración

México, D. F., Junio de 2015

Dedicatoria

Dedico este trabajo a las dos mujeres que llenan mi vida de felicidad: mi hija y mi esposa.

Ariadna Sofía Serralde:

Bendices mi vida a diario con tus besos, abrazos, juegos y tu compañía. Me has recordado muchas cosas importantes que los adultos solemos olvidar de la vida, como lo grandioso que es ser un príncipe, o lo grato que es dormir cinco minutos más en las mañanas, abrazado a ti. Espero estar haciendo lo correcto para contribuir a que seas una mujer feliz.

Mi amor, te dedico este trabajo porque tu esfuerzo y dedicación en todo lo que haces me ha inspirado para esforzarme también. Lejos de pretender enseñarte algo, quiero aprender de ti. Te amo.

Yuleimi Ayala:

Esposa mía, compañera de vida. Hemos vivido muchas aventuras, hemos reído innumerables ocasiones, hemos aprendido cosas nuevas, hemos disfrutado de interminables pláticas que nos han hecho conocernos a la perfección, hemos crecido juntos profesionalmente y como personas.

Este es un logro más que vivimos juntos. Te dedico este trabajo porque eres la cómplice de todas mis locuras. Esto no hubiera sido posible sin ti. Mis más sinceras gracias por ser parte inseparable de mi vida. Te amo.

Agradecimientos

A Dios, por los “talentos” que ha depositado en mi persona. Declaro que son suyos y que solo soy el trabajador. Gracias por tantas bendiciones que he recibido, en especial la vida y mi familia.

A mis padres, por su apoyo incondicional, por haberme inculcado la disciplina y el respeto. Papá, mamá, ¡los amo!

A mis tutores, Dr. Carlos Francisco Méndez Cruz y Dra. Azucena Montes Rendón por su guía, apoyo y paciencia. Gracias por mostrarme este mundo maravilloso de la ciencia e introducirme en él.

A Carlos Méndez, la persona. Gracias, Carlos, por la confianza que depositaste en mi persona, por compartirme tus conocimientos y experiencias. He aprendido mucho, sobre diversas cosas, gracias a ti.

Al Dr. Gerardo Sierra Martínez por darme la oportunidad de ser parte de su excelente equipo de trabajo.

Al Grupo de Ingeniería Lingüística de la UNAM por brindarme el espacio y los recursos necesarios para la realización del presente trabajo de investigación.

Al Consejo Nacional de Ciencia y Tecnología, que a través de los proyectos *Detección y medición automática de similitud textual* (Clave de registro: 178248) y *Caracterización de huellas textuales para análisis forense* (Clave de registro: 215179) me otorgó una beca que me permitió realizar mis estudios de maestría y realizar este trabajo de investigación.

A mi familia natural y política, por sus muestras de cariño. Sé que puedo contar con todos ustedes.

A mis compañeros del GIL que compartieron conmigo su juventud y alegría. Admiro el enorme potencial que cada uno de ustedes posee.

Contenido

Índice de tablas

Índice de figuras

Introducción

Vivimos en la era del conocimiento (Martínez Sánchez & Corrales Estrada, 2011) y hemos sido testigos de cómo este ha facilitado la transformación del panorama económico mundial. Ejemplo de ello es la apuesta que hacen algunos países asiáticos al dejar su liderazgo manufacturero para desarrollar una economía basada en el conocimiento.

Hoy en día el conocimiento es reconocido como recurso protagónico en la generación de valor para las organizaciones. Dicho reconocimiento inició con los trabajos de Thomas Stewart y Ruckdeschel (1998). En ellos se adopta el término *capital intelectual* para referirse al cúmulo de conocimiento inmerso en una organización.

Trabajos posteriores como el de Ordoñez De Pablos (2001) sostienen que el conocimiento que posee una organización es un capital intangible que tiene el potencial de servir como recurso estratégico para diferenciarla de las demás y proporcionarle una ventaja competitiva. En esa misma línea, Vorakulpipat & Rezgui (2008) reconocen que dicho conocimiento es un generador del capital intelectual de la organización debido a que este tiene la posibilidad de aplicarse para facilitar y mejorar el aprendizaje organizacional, la innovación, las capacidades, competencias y experiencia de los empleados.

Conviene mencionar que el conocimiento de una organización puede estar presente de manera explícita o tácita (Nonaka, 1994). En el primer caso, la información está plasmada en documentos, ya sea físicos o electrónicos, que pueden compartirse y ser recuperados fácilmente. Para este tipo de documentos, la extracción de información relevante y oportuna representa uno de los problemas más grandes. En el segundo caso, el conocimiento se encuentra de manera implícita en la mente de los integrantes de la organización a manera de experiencia, interpretación de las cosas y procesos de reflexión. En este tipo de conocimiento, el reto radica en hacer su extracción desde la mente del individuo.

En consecuencia, es importante encontrar la forma de administrar este elemento intangible como un recurso más de una organización y para tal fin, la *administración del conocimiento* ha sido fructífera, pues facilita la identificación, creación, preservación, comunicación, y aplicación del conocimiento de las organizaciones para generarles valor.

De manera general, se puede decir que la administración del conocimiento es una disciplina que consta de un conjunto de actividades generales que posibilitan la administración del conocimiento de una organización. Actualmente existe una gran cantidad de marcos de trabajo que proveen una visión general acerca de dichas actividades, también llamadas core, y ofrecen distintas propuestas en cuanto a cuáles y cuántas son las actividades necesarias para dicha administración. Sin embargo, es común encontrar las siguientes actividades: Identificar, crear, almacenar, compartir y aplicar; en ocasiones nombradas con distintos sinónimos (Heisig, 2009).

Cada actividad general o core de la administración del conocimiento ha aprovechado estrategias, métodos, actividades, procesos y desarrollos tecnológicos de comunicación y colaboración para alcanzar sus objetivos. Estas suelen llamarse *herramientas de la administración del conocimiento* y pueden ser, o no, basadas en tecnologías de la información (TI). Entre ellas se encuentran las herramientas de búsqueda, blogs, chats y otros elementos de colaboración y comunicación. El objetivo común de dichas herramientas es el incremento de la agilidad y facilidad de la interacción que permita un rápido, sostenido y continuo intercambio de conocimiento (Yang, Jing-Jun, & Chang-xiong, 2007).

El conjunto de las herramientas que facilitan la colaboración entre los integrantes de la organización y el acceso al conocimiento, la colección de documentos que contienen el conocimiento explícito, y el sistema que administra estos recursos suelen llamarse *repositorios de conocimiento* (Staab, 2001). Estos repositorios normalmente están integrados por intranets, cuentas de correo, manuales de procedimientos, noticias y otros documentos de la organización (Moldovan, 2001), además de blogs, chats y otros elementos de colaboración y comunicación.

Las herramientas tecnológicas nombradas con anterioridad han facilitado considerablemente la administración del conocimiento de la organización. Sin embargo, el hecho de descubrir, crear, almacenar, compartir y aplicar constantemente dicho conocimiento implica, entre otras cosas, la creación de una colección de documentos en diferentes formatos, sobre todo archivos de texto con información no estructurada (Ciravegna, 2001), que suele tener un tamaño considerable. Esto dificulta conseguir una organización adecuada de documentos que facilite la localización y entrega oportuna, económica y eficiente del conocimiento requerido.

Para mitigar este problema, es conveniente colocar los documentos en el repositorio según las áreas de conocimiento de la organización, que a su vez están determinadas por la misión y visión de la misma. Para tal efecto algunos marcos de trabajo sugieren la aplicación de algún elemento que revele la taxonomía del conocimiento y, por tanto, las áreas de conocimiento que las constituyen. Esta taxonomía del conocimiento se replica en la estructura de un sistema que administre los documentos (normalmente sistemas gestores de contenidos) y se ingresan los documentos en las áreas (secciones) correspondientes cuando estos se vayan generando.

Planteamiento del problema

La estrategia planteada anteriormente ha facilitado la organización del conocimiento de los repositorios. Sin embargo, la generación y actualización constante de conocimiento provoca que su inserción en la sección adecuada y documento correspondiente demande mucho tiempo y esfuerzo. De manera que hacer esto a través de un proceso manual en ocasiones es impráctico, limitado y hasta imposible.

El procesamiento del lenguaje natural (PLN), como parte de la inteligencia artificial y las tecnologías del lenguaje, es una disciplina que puede auxiliar el proceso de extracción y organización del conocimiento pues puede proveer herramientas de recuperación, extracción, resumen, traducción, presentación y generación de conocimiento, tal como lo expone el trabajo de Maybury (2001). Además, en el trabajo de Staab (2001) se exhibe la necesidad de aplicar herramientas de procesamiento de lenguaje humano para promover que toda la gente de la organización use los sistemas de administración del conocimiento y acceda a sus beneficios.

Hasta ahora, los mayores avances se han logrado en la creación de herramientas avanzadas de búsqueda y sistemas de localización de expertos (Maybury, 2001). Por lo que se percibe que existe una falta de herramientas que permitan la organización de repositorios de conocimiento de forma automática.

En consecuencia, se intuye que es posible diseñar una metodología que integre el uso de las tecnologías del lenguaje para procesar el conocimiento explicitado en las herramientas de la administración del conocimiento, para proveer un proceso semiautomático de organización de los repositorios. Lo anterior aporta a la administración del conocimiento en

relación a que una organización eficiente del conocimiento facilita la ejecución de sus actividades generales y el alcance de sus objetivos.

Así, en este trabajo se usarán técnicas de las tecnologías del lenguaje para extraer conocimiento y organizar semiautomáticamente el contenido de los repositorios de conocimiento, aprovechando la taxonomía utilizada en su construcción. Para esto, se usará como caso de estudio un grupo de investigación de la UNAM.

Sin embargo, un grupo de investigación posee características muy particulares de estructura, procesos internos, objetivos, misión y visión. Para este tipo de organizaciones, cuyo objetivo principal es la generación de conocimiento, el nuevo conocimiento resulta ser el valor producido. Por lo anterior, nos encontramos con dos tipos de conocimiento: Aquel que es utilizado para realizar los procesos de la organización (grupo de investigación) y aquel que surge como producto final.

Estas particularidades obligan a observar el grado de transparencia con el que se puede usar un marco de referencia en la implementación de la administración del conocimiento en un grupo de investigación, o en su defecto, identificar los elementos que deben adaptarse para dicha implementación.

Preguntas de investigación

El planteamiento del problema presentado anteriormente genera las siguientes preguntas de investigación:

1. Dado que en los repositorios de conocimiento se explicita continuamente el conocimiento de la organización mediante herramientas de colaboración y comunicación, ¿es posible extraer conocimiento de manera automática de un repositorio generado por la implementación de la administración del conocimiento, mediante el uso de las tecnologías del lenguaje?

2. Considerando que la organización manual de los repositorios de conocimiento requiere en muchas ocasiones de mucho esfuerzo y tiempo, ¿es posible organizar automáticamente un repositorio de conocimiento a través de las tecnologías del lenguaje, aprovechando la taxonomía del conocimiento que se utilizó en su construcción?

3. En caso de que sea posible organizar automáticamente un repositorio, ¿qué tecnologías del lenguaje pueden utilizarse para tal efecto?

4. Finalmente, dado que se tomará como caso de estudio el repositorio de un grupo de investigación, cabe preguntarse: ¿qué particularidades presenta la implementación de un modelo de administración de conocimiento en un grupo de investigación?

Objetivos

Los objetivos que se establecen en el presente trabajo para dar respuesta a las interrogantes planteadas son:

1. Diseñar una metodología para organizar automática o semiautomáticamente el contenido de repositorios de conocimiento mediante tecnologías del lenguaje.
2. Generar un repositorio de conocimiento mediante la implementación de la administración del conocimiento en un grupo de investigación.
3. Siguiendo la metodología diseñada, aplicar métodos y técnicas de las tecnologías el lenguaje para extraer conocimiento y organizarlo automática o semiautomáticamente en un repositorio.
4. Observar las particularidades de la implementación originadas por las características propias de la estructura y funcionamiento de un grupo de investigación.

Hipótesis

A manera de hipótesis se intuye que aplicando tecnologías del lenguaje es posible procesar el conocimiento explicitado en un repositorio producto de la implementación de la administración de conocimiento, específicamente en sus herramientas colaborativas, para proveer un proceso automático o semiautomático de organización de los repositorios.

Justificación

El presente trabajo de investigación contribuye a la informática administrativa y a la administración del conocimiento al exhibir la posibilidad de explotar y organizar automática

o semiautomáticamente los repositorios de conocimiento. Para tal efecto, se propondrá una metodología que aproveche las tecnologías del lenguaje para procesar automáticamente el conocimiento explicitado en las herramientas de la administración del conocimiento. Lo anterior aporta a la administración del conocimiento ya que facilita la ejecución de sus actividades generales y el alcance de sus objetivos a través de la organización eficiente del conocimiento.

Además, la implementación de la administración del conocimiento en el grupo de investigación que se ocupará como caso de estudio le proveerá las siguientes ventajas:

- Disminuir la curva de aprendizaje de sus nuevos integrantes en relación a los procesos administrativos de la organización.
- Preservar el conocimiento.
- Evitar la pérdida de conocimiento como resultado de la salida de alguno de sus integrantes.
- Mejorar y hacer reingeniería de los procesos internos.
- Facilitar a los integrantes de la organización el compartir y aplicar el conocimiento.

Viabilidad

El recurso principal para desarrollar la investigación propuesta es el repositorio de conocimiento que se genere tras la implementación de la administración del conocimiento en una organización. Por tanto, es indispensable asegurar la disponibilidad y apoyo de la organización que se ocupa como caso de estudio en la presente investigación.

El Grupo de Ingeniería Lingüística (GIL) de la Universidad Nacional Autónoma de México, a través de sus directivos, ha expuesto su interés y apoyo para la realización de las actividades y procesos necesarios para la realización de dicha implementación. Este apoyo, brindado desde el nivel jerárquico mayor, asegura la disponibilidad de los recursos humanos, materiales y de conocimiento necesarios para la realización de dicha implementación y las tareas adquiridas tras su realización. Los equipos de cómputo, software, espacios de trabajo, sistemas de telecomunicaciones y papelería serán provistos por esta organización.

El GIL ha producido un extenso conocimiento alrededor de las tecnologías del lenguaje y el respaldo ofrecido por la organización garantiza su acceso. Al hacerlo, nos provee del conocimiento necesario para realizar parte del trabajo de investigación.

Por otra parte, entre los diversos marcos de trabajo (*Frameworks*) para guiar el proceso de implementación de la administración del conocimiento, se ha elegido el uso del Framework de la Asian Productivity Organization (APO) debido al libre acceso a todos sus recursos (manuales, artículos, material para capacitación, publicaciones, etc.) y la existencia de literatura que documenta el éxito en la implementación de este marco de trabajo en otras organizaciones pequeñas. De esta manera, el costo de la implementación de la administración del conocimiento en el grupo de investigación resulta mínimo y su probabilidad de éxito es alta. Esto refuerza la probabilidad de contar con el repositorio de conocimiento que sirva como materia prima para desarrollar la presente investigación.

En consecuencia, los requerimientos técnicos, humanos, materiales, intelectuales, de tiempo y económicos no exhiben algún problema para el desarrollo de este trabajo.

Alcance

La metodología que se propone tiene la intención de ser aplicable a cualquier repositorio de conocimiento, haciendo adaptaciones al dominio en cuestión. Sin embargo, para fines de este trabajo, la experimentación se hace únicamente en el repositorio de conocimiento que se genera tras la implementación de la administración del conocimiento en el grupo de investigación que es utilizado como caso de estudio.

Se pretende diseñar una metodología que haga uso de las tecnologías del lenguaje para procesar el conocimiento explicitado en las herramientas de la administración del conocimiento, e identificar automáticamente las secciones del repositorio o los documentos probables donde este pudiera colocarse. Sin embargo, no se pretende sustituir el trabajo humano en la evaluación de la pertinencia de enviar dicho conocimiento a un espacio, a otro o a ambos, ni su inserción en este.

Las tecnologías del lenguaje cuentan con un extenso número de técnicas, métodos y desarrollos tecnológicos. A pesar de esto, la intención de este trabajo no es probar la aplicación de todas estas tecnologías, sino únicamente aquellas que son de utilidad en la metodología diseñada. No obstante, se intuye que es posible aplicar muchos de estos elementos para explotar y organizar repositorios de conocimiento.

Aunque en la investigación es necesario hacer una representación del conocimiento de la organización, no se pretende proponer algún modelo nuevo para hacerlo. En cambio se pretende aprovechar la taxonomía del conocimiento, que se genera tras la implementación de la administración del conocimiento, para definir la estructura del repositorio según la APO y

otros marcos de trabajo. Además, el modelo utilizado solo representará el conocimiento en el dominio específico de este trabajo con el nivel de detalle necesario para cumplir los objetivos planteados.

Estructura de la tesis

La presente tesis está dividida en seis capítulos. En el capítulo 1 se presenta una revisión del estado del arte sobre la administración del conocimiento. El capítulo 2 desarrolla la teoría básica de las tecnologías del lenguaje. El capítulo 3 se exhibe el proceso de implementación de la administración del conocimiento en el grupo de investigación ocupado como caso de estudio. En el capítulo 4 se muestra la aplicación de tecnologías del lenguaje para explotar el repositorio creado tras la implementación de la administración del conocimiento. El capítulo 5 muestra el diseño y propuesta de una metodología para la organización semiautomática del conocimiento que se crea en las herramientas de la administración del conocimiento. Finalmente, en el capítulo 6 se presentan las conclusiones del presente trabajo de investigación.

CAPÍTULO 1. Administración del conocimiento

En este capítulo se expondrá la teoría básica de la administración del conocimiento. Al principio se define qué es el conocimiento. Luego, se revisan los dos grandes tipos de conocimiento que existen en las organizaciones. Después, se analiza la importancia del conocimiento como elemento generador de valor y ventaja competitiva. En seguida, se presenta la administración del conocimiento, sus actividades generales, sus herramientas y el repositorio de conocimiento que se crea al implementarla. Finalmente se revisan algunas publicaciones científicas sobre la utilización de la administración del conocimiento en grupos de investigación.

1.1. El conocimiento

Es evidente el hecho de que el conocimiento juega un papel protagónico en la vida diaria. Nuestras actividades cotidianas son realizadas aplicando el conocimiento que hemos adquirido de fuentes externas o que hemos generado para resolver problemas. Todo este conocimiento es susceptible de modificarse pues, gracias a la observación, experimentación y relación de conceptos, se actualiza, refina, corrige o adquiere nuevo conocimiento. Esta característica es considerada por el sociólogo Luhmann, citado por Paz López, López Molina, & Solórzano Rodas (2011), al asegurar que el conocimiento es un *esquema cognitivo* que se considera al mismo tiempo verdadero y variable.

Lo anterior permite percibir que existe una concepción colectiva de lo que es el conocimiento. Sin embargo, hacer una definición que precise a qué se hace referencia con este nombre es muy complicado, pues existen diversas acepciones de lo que debe considerarse como tal. De hecho, es una tarea antigua que se remonta a la Grecia antigua, específicamente en la filosofía de Platón (Canals, 2003). Para este filósofo, el conocimiento es una *creencia cierta justificada*.

El problema principal con la definición anterior es que, el hecho de tener una creencia, obliga a tener conciencia de lo que se cree. Sin embargo, autores como Polanyi (1967) afirman que existe un tipo de conocimiento llamado tácito, en el que el individuo no es consciente de su existencia. Por ejemplo, el conocimiento que posee una persona sobre cómo usar una bicicleta es un elemento de esencia notoriamente distinta a una creencia.

Por su parte, la corriente de filosofía llamada empirista afirma que el conocimiento solo puede ser adquirido mediante la experiencia (Hume, 1957). Esta definición también tiene limitantes pues la reflexión e introspección también pueden generar nuevo conocimiento, como el caso del conocimiento contenido en teorías matemáticas o modelos abstractos como la teoría cuántica. Un trabajo posterior realizado por Biggam (2001) aporta una definición más amplia de conocimiento, al aceptar la definición del enfoque empirista y agregar que el conocimiento también puede ser el resultado de un pensamiento racional. En el mismo trabajo se asegura que, comúnmente, el conocimiento es el resultado de la unión de la experiencia sobre un suceso o elemento y el análisis racional del mismo.

Esta definición admite que es posible almacenar y transmitir el conocimiento a otros individuos sin la necesidad de que estos tengan que exponerse a la experiencia que otros ya procesaron, de manera que es posible pensar en una estrategia de intercambio y creación sostenida de conocimiento.

Por otro lado, conviene diferenciar al conocimiento de los datos o la información. Los datos son los elementos captados por los sentidos, mientras que la información es la consecuencia del procesamiento de los datos. Finalmente, el conocimiento es el entendimiento producido por el razonamiento. Además, Gianneto y Wheeler (2002) aseguran que el conocimiento integra varios elementos, entre los que se encuentran las creencias, valores, creatividad, juicio, habilidades, experiencia, teorías, reglas, relaciones, opiniones, conceptos y experiencias previas.

1.1.1. Conocimiento explícito y tácito, y creación de conocimiento

Nonaka (1994), en su artículo llamado A Dynamic Theory of Organizational Knowledge Creation, afirma que existen dos dimensiones en la dinámica de la creación del conocimiento en las organizaciones: epistemológica y ontológica.

La dimensión epistemológica describe que el conocimiento de una organización puede estar plasmado o codificado en documentos a través de algún sistema de lenguaje, o

puede quedar encerrado en la mente de las personas. En el primero de los casos, el conocimiento es llamado explícito y en el segundo, tácito.

El conocimiento explícito tiene la posibilidad de ser documentado y descrito para poder compartirse entre los integrantes de la organización a través de textos, sonidos, imágenes, videos, etc (Gianneto & Wheeler, 2002). Esta comunicación es indirecta, pues no es necesaria la interacción directa entre personas para poder realizarla. Los elementos en los que puede estar contenido el conocimiento explícito pueden ser muy variados, por ejemplo, manuales de procedimientos, reglamentos, oficios, memorándums, contratos, memorias técnicas, correo electrónico, wikis, blogs, archivos de computadora, contenido web e intranet, y demás documentos propios de la organización, ya sean físicos o electrónicos.

En cambio, el conocimiento tácito necesita una interacción más personal y directa para poder transmitirse, ya que este se encuentra inmerso en la mente de las personas de la organización en forma de experiencias personales, razonamiento, creencias, valores, actitudes, etc. que no pueden explicitarse tan fácilmente.

La segunda dimensión, la ontológica, hace referencia a la creación del conocimiento como un proceso continuo, dinámico y humano, de pensamiento en diferentes niveles: individual, grupal y organizacional. Al respecto, Canals (2003) advierte que el conocimiento colectivo no es la simple suma de los conocimientos individuales, pues la colectividad genera, además, conocimiento relacionado con la forma en que los conocimientos individuales se relacionan e interactúan. Por ejemplo, en cualquier equipo deportivo, el éxito no estará garantizado simplemente al agrupar varios jugadores expertos, se necesita, además que trabajen como un equipo, donde cada integrante conoce las características de los demás y las aprovecha para que el equipo tenga ventaja sobre otros.

Los trabajos de Nonaka (1994) aseguran que la creación del conocimiento se da mediante sucesivas transformaciones entre ambos tipos de conocimiento, a través de los siguientes procesos:

a. Externalización. Es la conversión de conocimiento tácito en conocimiento explícito mediante su síntesis y codificación a través de símbolos que permitan documentarlo. A esta codificación se le conoce también con el nombre de metáfora. La finalidad de que el conocimiento se codifique es que

éste pueda ser transmitido. Si el nivel de codificación y abstracción es bajo, es difícil que el conocimiento pueda difundirse (Canals, 2003).

b. Internalización. Consiste en convertir conocimiento explícito en tácito, mediante la asimilación del primero por parte de un individuo. Esto es equivalente a la definición de *aprendizaje*, utilizada en otras disciplinas.

También por la interacción del mismo tipo de conocimiento, específicamente a través de las siguientes formas:

c. Combinación. Consiste en tomar diferentes elementos de conocimiento explícito para combinarlos, ordenarlos, categorizarlos o agregarles nuevo conocimiento, frecuentemente, auxiliados de herramientas tecnológicas. El resultado de este proceso es la creación de nuevo conocimiento explícito.

d. Socialización. Mediante este proceso se crea nuevo conocimiento tácito a través de la interacción entre personas que comparten experiencias. La transmisión de dichas experiencias puede darse con o sin el uso del lenguaje, en este último caso, a través de la observación, imitación y práctica.

La siguiente imagen resume los cuatro procesos involucrados en la creación de conocimiento y su interacción con los dos tipos de conocimiento: tácito e implícito.

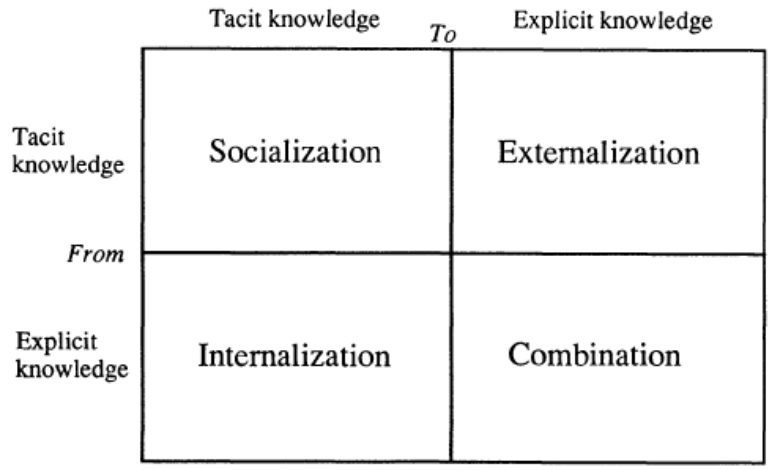


Figura 1-1 Procesos involucrados en la creación de conocimiento

Fuente: Nonaka, I. (1994), página 19.

1.1.2. El conocimiento como generador de valor

La identificación del conocimiento como un recurso estratégico importante se conoce desde la antigüedad, pues como señala Sánchez Medina, Melián González, y Hormiga Pérez (2007), los griegos y egipcios incrementaron su poder regional, entre otras cosas, gracias al conocimiento que codificaron y almacenaron en sus bibliotecas nacionales.

No obstante, en épocas recientes, específicamente en la segunda mitad del siglo XX, surgieron nuevas tecnologías que colocaron al conocimiento como elemento protagónico en la generación de innovación, aumento de la productividad y, en consecuencia, de competitividad. Esto se debe al hecho de que para poder aprovechar dichas tecnologías, se requiere mayormente de conocimiento que de materias primas, capital o mano de obra. Entre estas tecnologías se encuentran la energía nuclear, la biotecnología y, sobre todo, las tecnologías de la información y la comunicación. De hecho, para Paz, López y Solórzano (2011), la aparición de dichas tecnologías representa el inicio de la tercera revolución industrial.

Lo anterior promovió el surgimiento de nuevas estructuras sociales y económicas sustentadas en el conocimiento, y la conformación de una nueva sociedad caracterizada por la creación continua de conocimiento y el constante desarrollo de las facultades intelectuales (Benavides Velazco & Quintana García, 2003). A esta nueva sociedad Drucker (1995) la llamó *sociedad del conocimiento*, término que hoy en día es ampliamente utilizado, junto con otros como *sociedad de la información* o *sociedad red*, (Krüger, 2006) para referirse a esta realidad social que, en la perspectiva de Martínez Sánchez y Corrales Estrada (2011) constituye una redefinición de lo es significativo para la colectividad, pues, en esta realidad, el conocimiento se posiciona como el elemento principal generador de valor, riqueza y desarrollo.

Al respecto, los trabajos de la Asian Productivity Organization (2009) y de Martínez Sánchez y Corrales Estrada (2011) exhiben el hecho de que las economías de las naciones más desarrolladas están modificando sus políticas de desarrollo para aprovechar al conocimiento como potenciador de su crecimiento económico, de manera que se están transformando en economías basadas en el conocimiento. Los mismos trabajos enlistan

algunos nombres de países para ejemplificar este hecho, entre ellos Canadá, Japón y Singapur. Canals (2003) afirma que en las economías basadas en el conocimiento, las prácticas organizacionales de la economía industrial son insuficientes y, en algunos casos, inservibles.

Hoy en día, existe un reconocimiento generalizado sobre la importancia que el conocimiento tiene como recurso generador de valor para cualquier tipo de organización. Incluso, en los trabajos de Thomas Stewart y Ruckdeschel (1998) se adopta el término *capital intelectual* para referirse al cúmulo de conocimiento inmerso en una organización, a través de sus empleados, y que otorgan una ventaja competitiva. Los trabajos de Vorakulpipat y Rezgui (2008) aceptan y adoptan el término, y aseguran que el conocimiento que conforma dicho capital intelectual posee una gran relevancia gracias a que facilita y mejora el aprendizaje organizacional, la innovación y las capacidades, competencias y experiencia de los empleados.

Además, estos mismos autores (Vorakulpipat & Rezgui, 2008) advierten que por sí solo, el conocimiento no genera valor para las organizaciones, pues no basta con el simple intercambio del conocimiento existente entre los integrantes de una organización (como se consideraba en los inicios de la administración del conocimiento, tema que será tratado en el capítulo siguiente), sino una completa cultura de creación de conocimiento que instaure una continua interacción entre las personas, con la finalidad de crear nuevo conocimiento de manera sostenida. Esto último es facilitado por la existencia y promoción de una cultura del cambio.

Cabe mencionar que el término *capital intelectual* es mayormente utilizado en el área empresarial, mientras que en las disciplinas económicas se prefiere el término *activos de conocimiento*, en tanto que en el área contable es común encontrar el nombre *recursos intangibles* Lev (2001), citado en Sánchez Medina, Melián González, y Hormiga Pérez (2007).

1.1.3. Ventaja competitiva a través del conocimiento

Benavides Velazco y Quintana García (2003) advierten que las organizaciones actuales, inmersas en la sociedad del conocimiento, enfrentan las siguientes realidades:

- Globalización y liberación de los mercados
- Difusión de las tecnologías de la información y comunicación
- Sofisticación de la demanda
- Sofisticación de los competidores
- Sofisticación de los proveedores
- Fortalecimiento de los regímenes de propiedad intelectual
- Cuellos de botella en la efectividad empresarial
- Creciente importancia de las capacidades dinámicas
- Necesidad de entender las funciones cognitivas humanas

En esta realidad, es necesario encontrar elementos que le permitan a la organización tener una ventaja competitiva. El problema es que los elementos que pueden generar esta ventaja, como maquinaria, recursos humanos, marcas, etc., pueden ser adquiridos por los competidores con relativa facilidad (Benavides Velazco & Quintana García, 2003) a través de proveedores que los comercializan, de manera que estos no representan un diferenciador para la organización.

No obstante, existe un elemento que difícilmente puede comercializarse y, por tanto, adquirirse: el conocimiento. Consecuentemente, el conocimiento que posee una organización tiene el potencial de servir como recurso estratégico para diferenciar a una organización de las demás y proporcionarle lo que Porter (1995) llama *ventaja competitiva*. De hecho, el Centro de Investigación y Documentación sobre problemas de la Economía, el Empleo y las Cualificaciones Profesionales (2004) afirma que el conocimiento y la capacidad de crearlo y utilizarlo constituyen la principal fuente de ventaja competitiva sostenible para las empresas, regiones y sociedades completas.

Existe literatura que confirma esta idea, al exhibir casos de éxito de organizaciones de diversos giros (gubernamentales, empresas privadas, de asistencia social, centros educativos, etc.) que muestran que el conocimiento, con la adecuada administración, es un elemento generador de ventaja competitiva (APO, 2009, pág. 20). Esta adecuada administración hace referencia a la administración del conocimiento como disciplina. Autores como Ordoñez De Pablos (2001) señalan que la administración del conocimiento conforma un conjunto de estrategias que permiten gestionar al conocimiento como activo

estratégico. Aramburo (1999) asegura, además, que el fin último de la administración del conocimiento es la mejora y sostenimiento de la capacidad competitiva de las organizaciones.

Cabe mencionar que los trabajos del mismo autor afirman que es necesario conocer a detalle las clases de conocimiento que existen (generalmente se aceptan dos, explícito y tácito, aunque existen muchos autores que proponen más tipos de conocimiento) y la forma en que estas deben ser aprovechadas para generar ventaja competitiva. Aramburo (1999), asegura que no todo el conocimiento puede ser utilizado como elemento estratégico para adquirir y sostener ventaja competitiva, pues para que esto sea posible, el conocimiento debe tener las siguientes características:

a) Imperfecta replicabilidad o imitabilidad. Conocimiento que es único, difícilmente imitable por otros individuos u organizaciones. Ejemplos de este tipo de conocimiento son las experiencias, habilidades y destrezas de los miembros de una organización.

b) Imperfecta movilidad. Conocimiento que es mayormente valioso para la organización que para otros usos alternativos.

c) Durabilidad. Conocimiento que tiene una caducidad distante o que se perfecciona con el tiempo.

d) Imperfecta sustitubilidad. Cuanto menor sea la posibilidad de que exista un conocimiento que pueda remplazar al conocimiento que posee una organización, mayor ventaja competitiva se obtiene a través de este.

Además, el autor advierte que para generar ventaja competitiva es mayormente importante la capacidad de la organización para integrar el conocimiento en las actividades diarias, que el conocimiento mismo.

1.2. Administración del conocimiento

Para que una organización pueda producir el producto o servicio para la cual fue creada se requiere la intervención de múltiples recursos: humanos, tecnológicos, materiales y de conocimiento. El trabajo de Ganesh D. (2001) exhibe que estos recursos interactúan entre sí

para generar un valor para la organización. Por tanto, una adecuada administración de estos recursos deberá asegurar su disponibilidad en el momento y lugar en que se requiera y su uso eficiente y efectivo (Yang, Jing-Jun, & Chang-xiong, 2007).

A lo largo del tiempo ha surgido basta literatura que estudia las mejores formas de administrar los tres primeros elementos; sin embargo, el último parece tener un reto extra: la intangibilidad. De hecho, autores como Canals (2003) aseguran que el conocimiento no puede ser administrado debido a que es un elemento abstracto. Sin embargo, afirman que lo que sí es posible administrar son los activos de conocimiento. Dichos activos se crean a partir del conocimiento y pueden servir para generar nuevo conocimiento. Ejemplos de estos activos son las bases de datos, documentos, capacidades, procesos, rutinas, etc.

Es así que se hace evidente la necesidad de idear una estrategia que permita, entre otras cosas, almacenar y compartir el conocimiento generado por la misma organización, principalmente a manera de experiencias sobre el desarrollo cotidiano de las tareas y proyectos emprendidos, para que posteriormente se aplique por parte de los involucrados en la organización, con la finalidad de generar nuevas e innovadoras formas de trabajo cada vez más eficientes (Jennex, Olfman, & Addo, 2002).

Para tal efecto, surge la Administración del Conocimiento que, basado en Ganesh D. (2001) y la Asian Productivity Organization (2009), puede definirse como una disciplina que permite identificar, crear, almacenar, compartir y aplicar el conocimiento de la organización con la finalidad de que dicho conocimiento sirva como un generador de valor, innovación y ventaja competitiva.

De hecho, Martínez Sánchez y Corrales Estrada (2011) afirman que la administración del conocimiento surgió por la necesidad de identificar, valuar y capitalizar los activos intangibles y demás factores que intervienen en el proceso de creación de valor, en un mundo caracterizado por el uso generalizado de las TIC y la globalización. Para tal efecto, aprovechó elementos de disciplinas como la economía o la administración de información. Posteriormente, la aparición de modelos como el de Nonaka y Takeuchi facilitaron el desarrollo de un marco conceptual que permitió avances considerables en esta disciplina (Canals, 2003).

El desarrollo de la administración del conocimiento como disciplina ha tenido una constante evolución. Esta evolución se caracteriza por los diferentes enfoques sobre qué elemento debe absorber la atención y los esfuerzos para lograr una eficiente administración del conocimiento. Estos enfoques conforman lo que se conoce como generaciones de la administración del conocimiento. A continuación se detalla cada una de ellas basado en la Asian Productivity Organization (2009).

1.2.1. Generaciones de la administración del conocimiento

En los años noventa surge la primera generación de la administración del conocimiento. En esos años, los esfuerzos se centraron en desarrollar e integrar herramientas tecnológicas que apoyaran la transmisión del conocimiento. Esto se debió a que la aparición de dicha tecnología modificó las formas de trabajo y cambiaron la forma de ver las cosas.

Sin embargo, algunos años después se hizo evidente que la tecnología por sí sola era insuficiente, pues esta solo es un facilitador y acelerador, pero la creación, transmisión y aplicación del conocimiento requieren del trabajo individual y colectivo de los integrantes de la organización, por lo tanto se crearon comunidades colaborativas para fomentar dichas formas de trabajo. Esto representó la segunda generación de la administración del conocimiento.

En la tercera generación se observó que era necesario rediseñar los procesos de la organización con el fin de que el aprendizaje fuera un elemento integrado a los mismos. Dicho aprendizaje es generado como consecuencia de una eficiente administración del conocimiento. Esta etapa se sitúa históricamente alrededor del año 2000.

Posteriormente, las organizaciones y expertos en la materia se percataron que la gran capacidad de la administración del conocimiento como generador de valor obliga a considerarla como un componente estratégico de la organización. La integración de este elemento en la estrategia organizacional personificó la cuarta generación de la administración del conocimiento.

Finalmente, alrededor del año 2007 la implementación de la administración del conocimiento fue concebida como un elemento que debe trabajar no solo de manera local o

interna, sino como un trabajo interorganizacional para aprovechar los conocimientos de otras organizaciones en beneficio del conjunto involucrado. Esta concepción encarnó a la quinta generación de la administración del conocimiento.

La siguiente imagen muestra de manera gráfica lo anteriormente descrito.

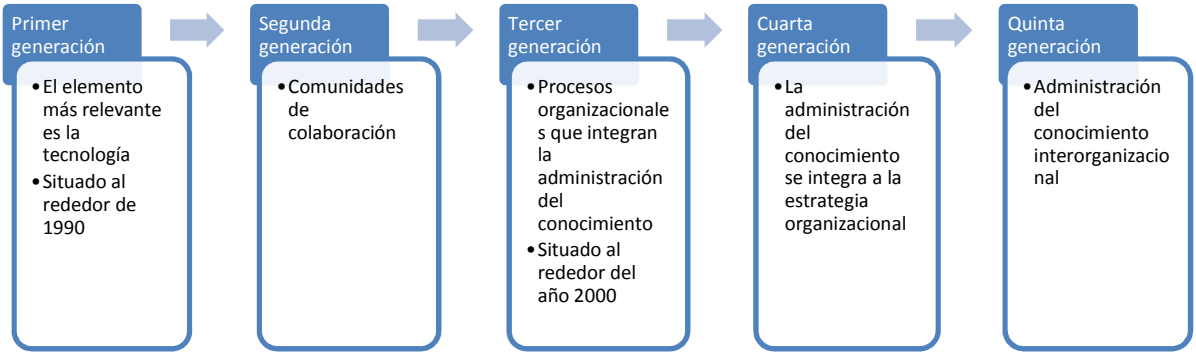


Figura 1-2 Generaciones de la administración del conocimiento

Fuente: Elaboración propia.

Como puede observarse, el nivel de importancia que la tecnología, las personas, los procesos y la estrategia tienen en la administración del conocimiento ha cambiado en las diferentes generaciones. Hoy en día es posible afirmar que las personas son el elemento fundamental en la administración del conocimiento, mientras que las tecnologías son una herramienta valiosa, siempre y cuando dichas tecnologías estén al servicio de una estrategia global con las personas como elemento central. No basta con la simple aplicación de las tecnologías (Canals, 2003).

El mismo autor afirma que, además de estos elementos, existe un catalizador para la creación de conocimiento en las organizaciones: El contexto. Ejemplo de ello son las comunidades de práctica, donde existe una clara y continua creación y transmisión de conocimiento. En estas comunidades el trabajo colaborativo es indispensable, por lo que Yang, Jing-Jun, & Chang-xiong (2007) advierten que es importante desarrollar estrategias, modelos y herramientas de comunicación y colaboración que faciliten la interacción.

Además, Canals (2003) observa que, dado que las personas son quienes deben explotar las herramientas tecnológicas que se implementen en la administración del conocimiento, se necesita que estas desarrollen habilidades y conocimientos sobre

recuperación de información, inteligencia competitiva, gestión documental, comunidades virtuales, usabilidad de la información, entre otras.

Por otro lado, Ordoñez De Pablos (2001) señala que la administración del conocimiento persigue cuatro objetivos principales:

1) Crear almacenes de conocimiento
2) Proporcionar acceso al conocimiento y facilitar su transferencia entre los miembros de la organización
3) Fomentar una cultura organizacional donde se privilegie la creación, transferencia y uso del conocimiento
4) Realizar auditorías del capital intelectual de la organización, como patentes, tecnologías, etc.

Los objetivos anteriores necesitan integrarse en una estrategia organizacional que permita aprovechar al conocimiento como fuente de valor para la organización y diferenciación con respecto a sus similares.

1.2.2. Actividades generales o core de la administración del conocimiento

La administración del conocimiento plantea un conjunto de actividades que deben realizarse para que el conocimiento de una organización pueda administrarse e insertarse en las tareas diarias para generar valor. Actualmente existe una gran cantidad de marcos de trabajo que proveen una visión general acerca de las actividades core y ofrecen distintas propuestas en cuanto a cuáles y cuántas son las actividades necesarias para dicha administración. Para facilitar una visión común sobre las actividades generales de la administración del conocimiento se han realizado esfuerzos por parte de organismos de estandarización europeos, australianos, británicos y alemanes.

Sin embargo, a pesar de las diferencias en la cantidad de actividades propuestas por los diferentes marcos de trabajo, es común encontrar las siguientes: Identificar, crear, almacenar, compartir y aplicar. En ocasiones estas actividades son nombradas con sinónimos. Al respecto, el trabajo de Heisig (2009), que compara 160 marcos de trabajo mediante

técnicas estadísticas, muestra que la media estadística en el número de actividades es de 5.1 con una desviación estándar de 1.7 y estas coinciden con las cinco nombradas anteriormente.

Basado en la APO (2009) y Ganesh D. (2001) se pueden definir las actividades generales de la siguiente manera:

1. Identificar. Esta actividad consiste en identificar el conocimiento requerido para ejecutar las tareas principales que facultan el logro de los objetivos de la organización y quién posee dicho conocimiento o dónde se encuentra. Esta actividad debe permitir conocer qué conocimiento está disponible y cuál no existe o es inaccesible.
2. Crear. Es la conversión del conocimiento existente y creación de nuevo conocimiento, a través de la combinación y reconfiguración de los conocimientos existentes para el desarrollo de nuevas ideas que sean útiles en la solución de problemas u optimización de actividades. Esta actividad se lleva a cabo en la interacción y la ejecución de los procesos y prácticas cotidianas de la organización.

   Esta actividad se desarrolla principalmente en dos formas:

   a) A nivel individual y de equipo. A través del entrenamiento, capacitación, aprendizaje por la experiencia de realizar las tareas, etc.
   b) A nivel departamental y organizacional. Mediante procesos de innovación dirigidos específicamente a la creación de conocimiento para nuevos productos, servicios, procesos, etc.

   La administración del conocimiento no solo debe permitir, sino promover la generación, revisión, actualización y mejoramiento del conocimiento por todos los integrantes de la organización (Canals, 2003).

3. Almacenar. Es la recolección y preservación del conocimiento de la organización a través de elementos y herramientas que faciliten su acceso eficiente. Para el manejo de conocimiento explícito esta actividad puede ser facilitada por herramientas tecnológicas. Sin embargo, para el conocimiento tácito es necesario encontrar formas de facilitar el acceso al poseedor de dicho conocimiento por parte de aquellos integrantes de la organización que lo requieran

4. Compartir. Consiste en el intercambio regular y sostenido del conocimiento entre los integrantes de la organización con la finalidad de fomentar el aprendizaje continuo para satisfacer las necesidades de la organización y lograr sus objetivos. Existen dos formas principales para compartir el conocimiento:
    a) Integrando bases de datos y distribuyendo documentos (enfoque de stock) y
    b) Transferencia persona a persona a través de la colaboración facilitada por técnicas como el *couching*, *mentoring*, etc. Al respecto, Canals (2003) manifiesta que frecuentemente la transmisión del conocimiento se dificulta, no por la dificultad técnica de hacerlo, sino por la resistencia y falta de disposición que las personas tienen para compartir su conocimiento con los demás integrantes de la organización. Para solventar este problema, la organización debe garantizar que el hecho de compartir el conocimiento beneficie a quien comparte el conocimiento, a quien lo recibe y a la organización misma. Además, la organización debe asegurarse que el flujo del conocimiento solo sea entre sus integrantes y no hacia el exterior, es decir, que no haya fugas de capital intelectual.

5. Aplicar. Consiste en la aplicación del conocimiento en los procesos del negocio con la finalidad de crearle valor. Esta actividad requiere de la creación e implementación de estrategias que motiven un pensamiento creativo que facilite la aplicación del conocimiento en la realización eficiente de las tareas y procesos necesarios para la creación de los productos y servicios de la organización.

1.2.3. Beneficios de la administración del conocimiento

Con base en la Asian Productivity Organization (2009) y Valerio (2002), los beneficios de la administración del conocimiento pueden observarse en tres niveles: individuales, organizacionales y sociales.

A nivel individual:

- Los involucrados perciben un incremento de sus conocimientos y habilidades
- Facilitación del trabajo

A nivel organizacional:

- Incremento de la productividad y la calidad de los productos y servicios ofrecidos
- Aumento en la satisfacción de sus usuarios
- Aseguramiento del flujo de conocimiento
- Se asegura la memoria organizacional
- Se promueve y acelera la innovación, tanto en los productos y servicios como en los procesos para producirlos, ya que como muestra la APO, el conocimiento es la base de la innovación.
- En consecuencia, la organización robustece sus fortalezas y aumenta su ventaja competitiva.

A nivel social:

- Las capacidades sociales se conforman por el conjunto de capacidades individuales y organizacionales, de manera que el incremento de estas últimas produce efectos positivos para la sociedad.

En resumen, el conjunto de beneficios individuales favorecen las capacidades organizacionales haciéndolas más competitivas y estas, a su vez, incrementan las capacidades de la sociedad.

1.2.4. Herramientas de la administración del conocimiento

La administración del conocimiento ha aprovechado estrategias, métodos, actividades, procesos y desarrollos tecnológicos de comunicación y colaboración para lograr los objetivos

de cada una de sus actividades generales. Este conjunto de elementos suelen llamarse herramientas de la administración del conocimiento y sirven para facilitar, mejorar y hacer más rápidos los procesos de generación, acceso, almacenamiento, codificación y transferencia del conocimiento de una organización (Ordoñez De Pablos, 2001).

Las herramientas de la administración del conocimiento pueden o no apoyarse en tecnologías de la información (TI). De hecho, en un principio se aprovecharon métodos muy simples como librerías físicas de documentos, oficinas para reuniones cara a cara, etc. (Supyuenyong & Islam, 2006) y a la fecha siguen utilizándose herramientas sin complejidad técnica mayor (Valerio, 2002). Sin embargo, hoy en día la tecnología es reconocida como un habilitador clave de las prácticas de la administración del conocimiento (Asian Productivity Organization, 2009), de manera que las herramientas de la administración del conocimiento han integrado una gran diversidad de tecnologías de la información y la comunicación que han permitido construir herramientas complejas que posibilitan interactuar eficientemente con el conocimiento de la organización y facilitar la colaboración entre sus miembros (Grau, 2003).

Al respecto, autores como Tyndale (2002) aseguran que es necesario incorporar varios tipos de tecnología para lograr una administración del conocimiento eficiente, y advierte que las tecnologías integradas en la administración del conocimiento deben ser revisadas constantemente, pues al paso del tiempo se vuelven obsoletas. Además, afirma que las organizaciones no explotan completamente el potencial de las tecnologías que poseen.

El mismo autor asegura que las herramientas de la administración del conocimiento que usan tecnologías de la información pueden clasificarse dentro de las seis categorías siguientes:

1) Sistemas administradores de documentos
2) Sistemas administradores de información
3) Sistemas de indexado y búsqueda
4) Sistemas expertos
5) Sistemas de comunicación y colaboración
6) Sistemas administradores de activos intelectuales

Además, Valerio (2002) observa que las herramientas existentes se aplican en la administración del conocimiento principalmente en las siguientes tareas:

a) Inteligencia empresarial
b) Aprendizaje organizacional
c) Diseño y administración de procesos
d) Identificación, desarrollo y administración de competencias
e) Administración de la experiencia

Como se ha mencionado, la finalidad de estas herramientas no es administrar el conocimiento por sí mismas, sino hacer posible o facilitar los procesos de la administración del conocimiento (Tyndale, 2002). Por tanto, cada herramienta puede auxiliar a varias actividades generales de la administración del conocimiento y, a su vez, cada actividad general puede estar apoyada por varias herramientas. En la siguiente tabla, construida con base en Supyuenyong e Islam (2006) y Asian Productivity Organization (2010), se aprecia este hecho. En ella se muestran algunos ejemplos de herramientas de la administración del conocimiento y las actividades que estas apoyan.

	Identificar	Crear	Almacenar	Compartir	Usar
Herramientas basadas en TI	Comunidades de práctica en línea Herramientas de búsqueda avanzada Localizador de Expertos Ambientes Colaborativos Virtuales	Comunidades de práctica en línea e-learning Bases de conocimiento (Wikis, etc.) Blogs Herramientas de búsqueda avanzada Localizador de Expertos Ambientes Colaborativos Virtuales Minería de datos	Comunidades de práctica en línea Bibliotecas de documentos Bases de conocimiento (Wikis, etc.) Blogs Localizador de Expertos Ambientes Colaborativos Virtuales Taxonomía del conocimiento Bases de datos Sistemas administradores de documentos	Comunidades de práctica en línea Bibliotecas de documentos Bases de conocimiento (Wikis, etc.) Blogs Redes sociales Localizador de Expertos Ambientes Colaborativos Virtuales Sistemas administradores de documentos	Comunidades de práctica en línea Sistemas expertos Sistemas de comunicación y colaboración Inteligencia artificial Bibliotecas de documentos Sistemas de modelado (virtualización) Bases de conocimiento (Wikis, etc.) Blogs Herramientas de búsqueda avanzada Localizador de Expertos Ambientes Colaborativos Virtuales Sistemas administradores de documentos

Tabla 1-1 Ejemplos de herramientas basadas en TI.

Fuente: Elaboración propia con base en Asian Productivity Organization (2010).

	Identificar	Crear	Almacenar	Compartir	Usar
Herramientas no basadas en TI	Cafés del Conocimiento Comunidades de práctica Mentoría	Lluvia de ideas Revisiones después de la acción Espacios colaborativos físicos Cafés del Conocimiento Comunidades de práctica Mentoría	Revisiones de aprendizaje Revisiones después de la acción Cafés del Conocimiento Comunidades de práctica	Asistencia de pares Revisiones de aprendizaje Revisiones después de la acción Comunidades de práctica Espacios colaborativos físicos Cafés del Conocimiento Mentoría	Asistencia de pares Espacios colaborativos físicos Cafés del Conocimiento Comunidades de práctica Plan de competencias Mentoría

Tabla 1-2 Ejemplos de herramientas no basadas en TI.

Fuente: Elaboración propia con base en Asian Productivity Organization (2010).

Cabe mencionar que Supyuenyong e Islam (2006) advierten que existen autores que aseguran que las tecnologías de la información no solo facilitan los procesos de la administración del conocimiento, sino que son capaces de crear nuevo conocimiento. Tal es el caso de la minería de datos que permite descubrir conocimiento en bases de datos u otras fuentes de datos estructurados; así como el de la minería de textos, o en general las tecnología del lenguaje, para descubrir conocimiento en colecciones de documentos. Esta posibilidad que tienen algunas tecnologías será abordada en capítulos posteriores para exhibir la posibilidad de aplicar las tecnologías del lenguaje para explotar el conocimiento de la organización de manera más eficiente y productiva.

Finalmente, Valerio (2002) asegura que es importante seleccionar las herramientas adecuadas para cada implementación de la administración del conocimiento en específico, pero sin perder de vista el hecho de que las herramientas por sí mismas no resuelven nada.

Además, advierte que el éxito de las herramientas depende en mayor grado de las personas, los procesos y la cultura organizacional, que de los factores técnicos.

1.2.4.1. Taxonomía del conocimiento

La creación de documentos para almacenar conocimiento explícito es una tarea cotidiana en la administración del conocimiento. Para poder utilizar el conocimiento contenido en ellos se aprovechan las herramientas de búsqueda, almacenamiento y extracción existentes (Grau, 2003). Sin embargo, existen elementos que pueden hacer más eficiente el trabajo de estos elementos, uno de ellos es una taxonomía.

El término taxonomía es utilizado comúnmente en la biología para referirse a la clasificación de organismos. Sin embargo, en la administración del conocimiento se aprovecha para organizar el conocimiento de una organización, pues se sabe que una adecuada organización del conocimiento disminuye el tiempo y esfuerzo necesarios para que un integrante de una organización pueda acceder al conocimiento que requiera.

Con base en el trabajo de la Asian Productivity Organization (2010), se puede decir que la taxonomía del conocimiento es una herramienta de la administración del conocimiento que permite encontrar una estructura adecuada para organizar los documentos, información y demás activos de conocimiento de una organización. Esta estructura tiene una forma jerárquica notablemente intuitiva que sirve, además, para facilitar la navegación, entrega y almacenamiento del conocimiento necesario.

La siguiente imagen muestra un ejemplo de una taxonomía del conocimiento.

Dimension (Level 1)	Management Domain			
Description of Dimension	This a repository of all knowledge assets relevant to management			
Level 2	Level 3	Level 4	Level 5	Comments
Knowledge Management				Contains all reference materials relating to KM
	General Concepts			
		KM Frameworks		Contains all reference materials relating to KM Frameworks
		Implementation Approaches		Contains all reference materials relating to Implementation Approaches
	Tools & Techniques			
		COP		Contains all reference materials relating to COP
		Storytelling		Contains all reference materials relating to Storytelling

Figura 1-3 Ejemplo de taxonomía del conocimiento.

Fuente: Asian Productivity Organization (2010). Página 40

La taxonomía del conocimiento hace frente a la necesidad de organizar estratégicamente al conocimiento en áreas, para tal efecto, la elaboración de dicha taxonomía comienza con la revisión de la misión, visión y objetivos de la organización. Una vez hecho esto, se identifican las áreas clave de conocimiento que permiten que esos objetivos, misión y visión sean alcanzados, es decir, los tipos o áreas de conocimiento que la organización necesita para desarrollar sus procesos y alcanzar sus objetivos y, en consecuencia, llegar a ser lo que se ha propuesto en su visión.

Luego, se identifican las personas, comunidades de práctica y redes de conocimiento que poseen el conocimiento de cada área localizada. Finalmente, los procesos, prácticas, políticas, métodos, y demás conocimiento, son organizados en las áreas de conocimiento encontradas.

Los mismos trabajos de la Asian Productivity Organization (2010) advierten que se debe tener cuidado en el nivel de detalle de la taxonomía del conocimiento que se elabore, ya que si la estructura de esta es demasiado detallada, el contenido es difícil de entregar. Por

el contrario, si dicha estructura tiene un nivel de detalle muy bajo, la taxonomía no será funcional.

1.2.5. Repositorios de conocimiento

Como se ha mencionado en secciones anteriores, las herramientas de la administración del conocimiento incrementan la agilidad y facilidad de la interacción y, en consecuencia, permiten un ágil, sostenido y continuo intercambio de conocimiento en las organizaciones (Yang, Jing-Jun, & Chang-xiong, 2007). Por tanto, conviene que estas herramientas estén accesibles y disponibles. En el caso de las herramientas basadas en TI es común encontrarlas integradas en un mismo lugar, normalmente en un portal web. Este portal debe proporcionar un acceso seguro, adaptable y personalizable a la información de la organización, normalmente desde distintos orígenes y en diferentes formatos (Loebbecke & Crowston, 2012).

Por otro lado, el hecho de descubrir, crear, almacenar, compartir y aplicar constantemente el conocimiento de la organización implica, entre otras cosas, la creación de una colección de documentos en diferentes formatos, sobre todo archivos de texto con información no estructurada (Ciravegna, 2001), que plasme el conocimiento explícito de la organización.

El conjunto de estas colecciones de documentos, las herramientas de la administración del conocimiento y el sistema que administra estos recursos suelen llamarse *repositorio de conocimiento* (Staab, 2001). Estos repositorios normalmente están integrados por intranets, cuentas de correo, manuales de procedimientos, noticias y otros documentos de la organización (Moldovan, 2001), además de sistemas de búsqueda y administración de contenidos, blogs, chats y otros elementos de colaboración y comunicación.

Loebbecke y Crowston (2012) aseguran que la finalidad de estos repositorios de conocimiento, también llamados *portales de conocimiento*, es soportar y estimular la transferencia, almacenamiento, entrega, creación, integración y aplicación del conocimiento a través de provisión de un simple punto de acceso al conocimiento disponible en la organización. Para esto, los repositorios de conocimiento deben sincronizar las diversas herramientas que procesen el conocimiento de la organización (diferentes sistemas usando

diferentes formatos) para proporcionar dicho acceso unificado a la información diseminada (Benbya, Passiante, & Aissa, 2004).

Estos últimos autores, después de comparar diversos repositorios de conocimiento, señalan que, además de las características que ya se mencionaron, estos deben ser:

a) Personalizables: El sistema debe ser capaz de entregar información relevante para cada usuario, según el trabajo que desempeñe en la organización.
b) Escalables: Debe ser posible la expansión de sus capacidades.
c) Seguros: Deben integrar herramientas que permitan controlar los accesos a los activos de conocimiento.
d) Colaborativos: Incluir herramientas colaborativas y de comunicación para facilitar y promover la creación de conocimiento mediante la socialización y su distribución.
e) Facilidad de publicación. Se debe permitir la publicación ágil y sencilla de documentos de diferentes formatos.

Además, deben integrar:

a) Taxonomías. Los documentos contenidos en los repositorios de conocimiento deben colocarse en categorías que mantengan una estructura jerárquica para facilitar la navegación y acceso al conocimiento. Para hacerlo es común utilizar taxonomías.
b) Sistemas de búsqueda. Los repositorios de conocimiento deben tener herramientas que permitan localizar el conocimiento requerido sin importar que este conocimiento se encuentre disperso en diferentes lugares del repositorio.

Sin embargo Benbya, Passiante, y Aissa (2004) y Loebbecke y Crowston (2012) advierten que muchas organizaciones enfrentan problemas para obtener beneficios notorios de dichos repositorios, limitándose comúnmente a resultados modestos. Estos últimos autores afirman que para evitar este problema, los repositorios de conocimiento deben poseer los tres elementos siguientes:

a) **Suficiente contribución:** Los portales o repositorios de conocimiento necesitan de la contribución o aporte de conocimiento por parte de los integrantes de la organización. Por lo tanto, es necesario que estos comprendan que el conocimiento de la

organización no debe ser un valor personal, sino de la organización, y que el compartir su conocimiento con los demás integrantes de la organización beneficia a toda la organización, incluyendo a quien lo comparte. El que cada integrante comprenda la importancia de su participación facilita la contribución del conocimiento de este al repositorio.

b) **Cultura organizacional favorable:** La cultura organizacional puede impactar de forma positiva o negativa en el uso de los repositorios de conocimiento como medio para compartir el conocimiento. Por ejemplo, una cultura organizacional donde se fomenta la competitividad individual provoca que los individuos atesoren sus conocimientos para explotarlos ellos mismos. Por lo contrario, una cultura que fomenta el apoyo entre los miembros de una organización facilita el intercambio de conocimiento. En consecuencia, para que un repositorio de conocimiento genere valor para la organización debe existir una cultura organizacional adecuada.
c) **Integración del conocimiento de la organización:** Un repositorio de conocimiento debe proporcionar un acceso centralizado a los diferentes orígenes de activos de conocimiento de la organización. Para lograrlo, debe diseñar, crear, o aprovechar mecanismos que permitan integrar dichas fuentes de conocimiento. En la medida en que esta integración sea mejor realizada, el repositorio de conocimiento es más eficiente para generar valor para la organización.

1.2.5.1. Organización de los repositorios de conocimiento

Loebbecke y Crowston (2012) aseguran que la organización del conocimiento constituye el componente más importante de los repositorios de conocimiento, pues, como Garud y Kumaraswamy (2005) afirman, dichos repositorios tienden a generar una sobrecarga de información por la facilidad que estos otorgan para acumular conocimiento en forma digitalizada, provocando que sea difícil aprovechar este conocimiento. En consecuencia, estos autores afirman que es importante realizar una categorización del conocimiento contenido en los repositorios para facilitar su recuperación.

Trabajos como los Benbya, Passiante, y Aissa (2004), Loebbecke y Crowston (2012), y Garud y Kumaraswamy (2005) muestran que la taxonomía, descrita en una sección anterior, es frecuentemente utilizada para realizar dicha categorización o clasificación. Incluso,

Benbya, Passiante, y Aissa (2004) usan los términos *esquemas de categorización o esquemas de clasificación* para referirse a las taxonomías, y aseguran que dicha taxonomía es un elemento básico y comúnmente usado en la organización de los portales o repositorios de conocimiento.

Garud y Kumaraswamy (2005) expresan la importancia del uso de la taxonomía para organizar el conocimiento de los repositorios al asegurar que esta no es solo una manera para categorizar dicho conocimiento, sino una estrategia para unificar diferentes grupos de conocimiento.

La taxonomía aplicada en los repositorios de conocimiento permite concentrar activos de conocimiento en grupos que representan temas diferentes. Cada tema, a su vez, puede dividirse en subtemas y cada uno de ellos forma un subgrupo de elementos. Con esto se crea una estructura jerárquica que facilita la búsqueda, navegación, control de acceso y entrega del conocimiento requerido (Asian Productivity Organization, 2009), (Benbya, Passiante, & Aissa, 2004).

Por otra parte, trabajos como los de Dias (2001), Maedche, Motik, Stojanovic, Studer, y Volz (2003), y Loebbecke y Crowston (2012) muestran que también es posible aprovechar elementos como los metadatos, los tesauros y las ontologías para el mismo fin y lograr una organización más eficiente, al mismo tiempo que se robustecen las capacidades de los sistemas de búsqueda integrados en los repositorios. Al respecto, Maedche, Motik, Stojanovic, Studer, y Volz (2003) aseguran que la siguiente generación de sistemas de administración del conocimiento dependerá en gran medida de las ontologías. Además, Staab (2001) muestra los avances y las discusiones del uso de estas tecnologías en la administración del conocimiento.

Los elementos anteriores permiten entregar a un usuario información extra sobre el activo de conocimiento que se consulta, permitiendo relacionarlo con otros activos de conocimiento facilitando la navegación dentro de los repositorios de conocimiento y permitiendo las búsquedas semánticas (Loebbecke & Crowston, 2012). Lo anterior aporta a la localización y entrega del conocimiento.

1.3. Administración del conocimiento en grupos de investigación

Canals (2003) expone que las actividades que conforman el proceso de administración del conocimiento (generación, almacenamiento, transmisión y aplicación de conocimiento, además de la revisión del conocimiento generado por un individuo por parte de sus pares o "peer review") ya lo hacía la ciencia desde antes de desarrollarse la administración del conocimiento como disciplina. Además, asegura que, considerando los avances que esta ha logrado, la ciencia es el sistema de administración del conocimiento más efectivo en la historia de la humanidad.

El mismo autor señala que la universidad es una entidad donde la administración del conocimiento puede ser muy fructífera, pues en ella existen grupos de investigación que tienen la misión de producir conocimiento, generar tecnología y desarrollar el talento de sus integrantes para crear recursos humanos profesionales en un área de estudio. Linlin y Hui (2008) indican que, comúnmente, el objetivo final que este tipo de organizaciones persigue es la innovación, la creación de tecnología, el desarrollo del talento humano y la creación de conocimiento, principalmente a través de la conversión del conocimiento individual tácito en conocimiento explícito grupal.

Por su parte, los resultados del trabajo de Yang, Chang-xiong, & Xue-mei (2006) demuestran que existen características similares entre las organizaciones con fines comerciales y los equipos de investigación y que, por lo tanto, es posible y adecuado implementar la administración del conocimiento con la reserva de tomar en consideración las particularidades de estos equipos de investigación al establecer el modelo a implementar.

Una de esas particularidades es que en los marcos de trabajo de la administración del conocimiento por lo general se identifica al conocimiento como como un insumo de los procesos de la organización, es decir, un elemento que sirve para desarrollar eficientemente los procesos productivos de una organización. En cambio, en los grupos de investigación, el conocimiento aparece en la mayoría de las ocasiones como insumo del proceso de investigación y como producto del mismo proceso. Este hecho también es visible en organizaciones de tipo empresarial, pero en un grado notablemente menor, tal como lo exhibe

el trabajo de Heisig (2009). La siguiente imagen pertenece a dicho trabajo y muestra esta doble categorización del conocimiento.

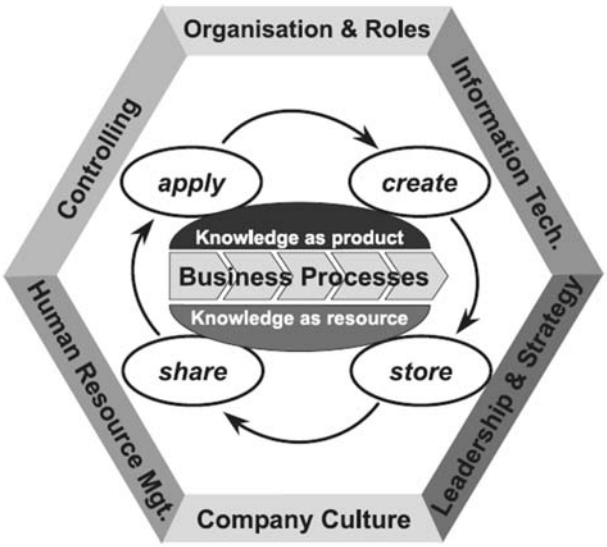


Figura 1-4 Marco de trabajo de Heisig para la administración del conocimiento

Fuente: Heisig (2009)

Sin embargo, Rivera (2000) exhibe que, contrario al protagonismo que la universidad y sus grupos de investigación debieran tener en cuanto a experiencias sobre la administración del conocimiento aplicada en sus tareas cotidianas, esta disciplina es investigada y enseñada a través de cursos académicos, pero no está siendo aplicada en ella misma para mejorar su desempeño. Además, indica que la participación de las instituciones educativas en las publicaciones sobre administración de conocimiento que se encuentran disponibles en la World Wide Web (WWW) es mínima, con un aproximado del 11% del total de publicaciones. Así mismo, afirma que las universidades del siglo XXI deberán reconstruirse sobre la teoría de la administración del conocimiento que en ellas mismas se gesta para adecuarse a los retos actuales.

Con relación a los retos que enfrenan hoy en día las universidades y grupos de investigación, Yang, Chang-xiong, Lei, Li-yan, y Ying (2008) advierten que el conocimiento individual de cada investigador debe comunicarse a la colectividad para que sea ella la propietaria y se obtenga un beneficio social. Por su parte, Yang, Chang-xiong, y Xue-mei (2006) advierten que en la actualidad los desafíos a los que se enfrenta la ciencia requieren que se generen equipos de trabajo multidisciplinarios que permitan la integración de los

conocimientos y tecnologías de diversas disciplinas en un trabajo colaborativo, de manera que el hecho de compartir conocimiento figura como el mecanismo principal para generar innovación y nuevo conocimiento Yang, Chang-xiong, Lei, Li-yan, & Ying (2008).

Además, es habitual que los equipos multidisciplinarios integren a expertos de diferentes regiones geográficas, así que se deben buscar mecanismos para facilitar el trabajo colaborativo a distancia para aprovechar las capacidades de cada uno de esos expertos (Yang, Jing-Jun, & Chang-xiong, 2007). En consecuencia, es necesario usar un modelo de trabajo que facilite los procesos de intercambio de conocimiento y el trabajo colaborativo. La implementación de un modelo de administración del conocimiento se observa oportuna para tal efecto pues ha demostrado su efectividad en diversos tipos de organizaciones, como gobierno e industrias productoras de bienes y servicios (Chandra & Khanijo, 2009).

Canals (2003) afirma que la administración del conocimiento puede aplicarse en los grupos de investigación para enfrentar dichos retos pues esta puede aprovecharse para facilitar las siguientes actividades:

a) Gestión de proyectos
b) Organización de congresos
c) Obtención de fondos
d) Facilitar el contacto y colaboración entre investigadores
e) Trabajo interdisciplinario
f) Creación de bases de datos sobre los intereses de investigación de personas, grupos de investigación y redes
g) Transmisión de resultados obtenidos en los trabajos de investigación
h) Obtención y comunicación de información sobre procedimientos, políticas y servicios que afectan la tarea del investigador

Además, advierte que, a pesar de que la investigación tiene principalmente esencia tácita, la explicitación del conocimiento relacionado con el proceso de investigación favorece al investigador en el ahorro de tiempo y esfuerzo.

Finalmente, Beesley & Cooper (2008) advierten que la comunidad científica dedicada a la investigación en administración del conocimiento debe ocupar esfuerzos en la

disminución de la brecha entre ella y la comunidad empresarial para facilitar el avance de esta disciplina.

CAPÍTULO 2. Tecnologías del lenguaje

En este capítulo se expondrá la definición de tecnologías del lenguaje y sus principales técnicas. En seguida, la exposición se centrará en explicar algunas de ellas, sus métodos y algoritmos. Finalmente se exhibirá la aplicación de las tecnologías del lenguaje en la administración del conocimiento.

2.1. Antecedentes

En secciones anteriores, se ha observado la importancia que el conocimiento tiene en nuestros días. Tal es el protagonismo que este elemento tiene para la sociedad actual, que se le conoce como Sociedad del conocimiento (Drucker & Centre Canadien de gestion, 1995). En este marco, se observa que la transmisión de este conocimiento es crucial, y es precisamente en dicha transmisión donde el lenguaje humano tiene un papel fundamental, pues como lo asegura Sierra (2009), la mejor herramienta para transmitir el conocimiento es el lenguaje.

Por otro lado, las tecnologías de la información y la comunicación han facilitado muchas tareas que antes requerían de enormes esfuerzos y recursos. El acceso y manipulación de conocimiento e información no es la excepción, tal como lo muestra Tyndale (2002). Según Sierra (2009), la enorme cantidad de información con que se cuenta hoy en día exhibe la función que tiene el lenguaje como medio de comunicación e interacción.

La unión de estos dos elementos, lenguaje y tecnologías de la información y comunicación, en una misma disciplina permite generar herramientas que facilitan la interacción entre personas y computadoras que tiene, entre otros objetivos, el manejo estratégico de la información (Sierra Martínez, 2009). Estas herramientas forman parte de lo que se conoce como tecnologías del lenguaje.

2.2. Definición

Las tecnologías del lenguaje son aquellas que se integran en programas informáticos para permitir el procesamiento del habla o de textos escritos (Llisterri, 2003). Para Sierra (2009), estas tecnologías del lenguaje incluyen conocimientos y medios necesarios para el tratamiento del lenguaje humano.

Los objetivos de las tecnologías del lenguaje, basado en Martí Antonín, y otros (2003), son las siguientes:

- Permitir la interacción entre máquinas y tecnología, y seres humanos, utilizando el lenguaje habitual para hacer más fácil dicha interacción.
- Facilitar el acceso a la información, superando restricciones como distancias.
- Permitir entornos multilingües. Es decir, el uso concurrente de diferentes lenguas.
- Interacción entre los diferentes tipos de comunicación (oral y escrita).
- Mejorar la comunicación en todas sus modalidades.
- Reconocer, comprender, interpretar y generar lenguaje humano.

Conviene mencionar que existen disciplinas relacionadas con las tecnologías del lenguaje. Una de ellas, muy relevante, es el Procesamiento del lenguaje natural (PLN) que como lo indica Sierra (2009) forma parte de la Inteligencia artificial y se encarga de la elaboración y aplicación de modelos teóricos y herramientas computacionales que permitan analizar, comprender, reproducir y manipular el lenguaje humano.

Existen otras dos disciplinas igualmente relacionadas con las tecnologías del lenguaje: la Lingüística computacional, que se encarga de modelar el lenguaje, y la Ingeniería Lingüística, que tiene la finalidad de crear programas informáticos que sean capaces de decodificar, comprender y generar lenguajes naturales

La siguiente imagen muestra gráficamente lo descrito en el párrafo anterior.

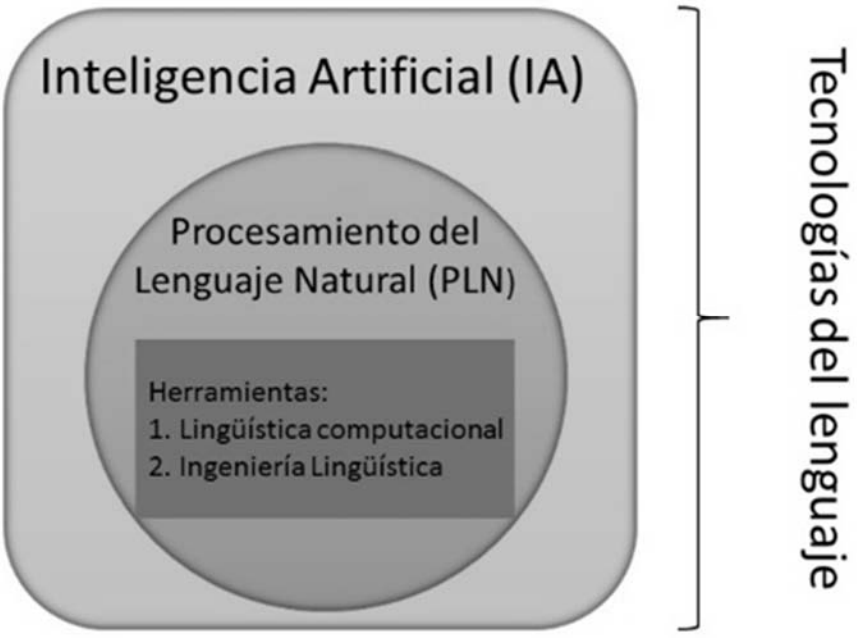


Figura 2-1 Tecnologías del lenguaje y su relación con otras disciplinas

Fuente: Elaboración propia

Conviene mencionar que existe otra disciplina estrechamente relacionada con las tecnologías del lenguaje llamada Minería de textos. Esta es un área interdisciplinaria que hace uso de la minería de datos y el aprendizaje de máquina, que a su vez integra estadística y computación (Hotho, Nürnberger, & Paab, 2005). La minería de textos tiene por objeto encontrar información y conocimiento relevante y previamente desconocido en documentos no estructurados, es decir, en texto en lenguaje natural (Gupta & Lehal, 2009).

2.3. Técnicas

Basado en los trabajos de Hotho, Nürnberger, y Paab (2005), Gupta & Lehal (2009), Weiss, Indurkhya, Zhang, y Damerau (2005) y Tan (1999), se presenta, a continuación, una lista con algunas de las técnicas desarrolladas en el procesamiento del lenguaje natural que pueden aplicarse para generar tecnologías del lenguaje.

a) Recuperación de información (*IR*, por sus siglas en inglés de *Information Retrieval*). Consiste en que dado un conjunto grande de documentos, se deben entregar aquellos que estén relacionados con un término o documento que nosotros indicamos como

entrada al sistema de recuperación de información. En otras palabras, el usuario entrega al sistema unas pistas o palabras clave sobre la información que necesita y el sistema entrega aquellos documentos que contienen tal información. En este sentido, Hotho, Nürnberger, & Paab (2005) advierten que los sistemas de recuperación de información buscan los documentos que contienen la información que responde a una pregunta del usuario, más no la respuesta en sí.

b) Extracción de información (*IE*, por sus siglas en inglés de *Information Extraction*). Consiste en extraer información específica de los documentos de texto.
c) Agrupamiento (*Clustering*). Es encontrar grupos de documentos que contengan contenidos similares.
d) Visualización de documentos. Radica en mostrar al usuario una representación gráfica de grupos de documentos con similitudes.
e) Seguimiento de temas. Consiste en predecir qué documentos pueden ser de interés a un usuario a través de la revisión de su perfil y el historial de los documentos vistos con anterioridad.
f) Resumen. Es reducir un texto en relación al detalle y la extensión, mientras que se mantienen las ideas principales y el significado del documento.
g) Consultas en lenguaje natural. Consiste en que un sistema pueda comprender una pregunta expresada en lenguaje humano y pueda procesarla para entregar la respuesta al usuario.

2.4. Clustering

El clustering, agrupamiento o conglomeración es un proceso que tiene como objetivo agrupar un conjunto de elementos dentro de algunos grupos o clusters, según la similitud de sus características (Gorunescu, 2011). Para tal efecto, cada elemento se representa como un vector que en cada dimensión contiene una característica del elemento (Weiss, Indurkhya, Zhang, & Damerau, 2005).

La intención es que los elementos de un cluster sean lo más similares entre ellos y lo más diferentes a los elementos de los otros clusters (Liu B. , 2007). Dicha similitud es determinada mediante un valor numérico descrito por alguna de varias funciones

matemáticas llamadas medidas de similitud o funciones de distancia. Entre ellas se encuentran las siguientes:

a. Minkowski.
b. Euclidiana.
c. Chebychev.
d. Tanimoto.
e. Pearson.
f. Mahalanobis.

2.4.1. Métodos

Según Liu (2007), existen dos tipos principales de clustering, el particional y el jerárquico. El primero únicamente genera grupos de elementos a un solo nivel, mientras que el segundo crea subgrupos dentro de los grupos principales de manera iterativa.

2.4.2. Algoritmos

Existen diversos algoritmos que realizan la tarea de clustering, entre ellos están el algoritmo EM y kMeans. A continuación se hará una breve descripción de ambos.

2.4.2.1. Algoritmo simple kMeans

Liu (2007) asegura que el algoritmo kMeans es ampliamente usado en las tareas de clustering particional, en gran parte, debido a su simplicidad y eficiencia. Basado en este autor y en Hartigan y Manchek (1979) se describe a continuación su funcionamiento.

Este algoritmo recibe como parámetro de entrada el número de clusters *k* que desean generarse. Con este dato, el algoritmo kMeans coloca ese número de puntos en un espacio de R dimensiones aleatoriamente (aunque permite indicarlo de manera manual). Estos puntos son llamados centroides y representan el centro de cada cluster.

Después, para cada elemento representado como un punto en el mismo espacio, se calcula la distancia hacia cada centroide y se asigna al más cercano. Una vez que se han asignado todos los elementos, los centroides son movidos a los centros geométricos delimitados por los elementos asignados a cada cluster. Esta operación es iterativa y termina cuando la localización de los centroides ya no cambia o cuando los elementos ya no son reasignados a otro cluster.

La principal desventaja de este algoritmo de clustering es que se necesita indicar el número de clusters que desean formarse y generalmente este dato no es conocido.

2.4.2.2. Algoritmo EM

Según lo expuesto en los trabajos de Moon (1996) y Weiss, Indurkhya, Zhang, y Damerau (2005), el algoritmo Esperanza–Maximización (*Expectation–Maximization*, EM) trabaja con bases matemáticas y estadísticas para encontrar información oculta o faltante en un conjunto de datos. Es un algoritmo que es utilizado en tareas de diversas áreas de conocimiento. La aplicación de este algoritmo de estimación estadística en tareas de aprendizaje automático corresponde al aprovechamiento de su capacidad para encontrar parámetros desconocidos.

De manera muy general, puede decirse que el algoritmo trabaja en dos etapas. En la primer etapa el algoritmo intuye y coloca un valor en el parámetro desconocido y en la segunda evalúa el resultado y modifica el valor del parámetro para maximizar la eficiencia del modelo (Redner & Walker, 1984).

El algoritmo EM aplicado a la tarea de clustering posibilita la búsqueda de los clusters de un conjunto de elementos sin tener la intuición del número de clusters que deben formarse. El algoritmo busca automáticamente el número de clusters que deben formarse e internamente va encontrando parámetros no visibles al usuario como el radio de las hiper esferas, la posición de los centroides, etc.

2.5. Extracción de información

Gupta y Lehal (2009) advierten que la extracción de información (*Information Extraction*) es el paso inicial para que las computadoras puedan procesar información no estructurada pues, como afirman Hotho, Nürnberger, y Paab (2005) esta consiste en extraer información

específica desde documentos de texto. La información extraída, habitualmente sirve para llenar plantillas o bases de datos que son más fácilmente procesables por las computadoras. Kodratoff (1999) asegura que esta información puede ser vista como información oculta en el texto.

Lo anterior es importante ya que, como afirman Gupta y Lehal (2009), un gran volumen de información del mundo se encuentra disponible únicamente en forma no estructurada, es decir, textos en lenguaje natural.

Basado en (Grishman, 2010) se indican a continuación las tareas comunes en la extracción de información.

- Extracción de nombres: Consiste en identificar nombres de personas, organizaciones, lugares, etcétera.
- Extracción de entidades: Es la identificación y vinculación de las frases que se refieren a los mismos objetos.
- Extracción de relación: Se trata de la identificación de pares de entidades en una relación semántica específica.
- Extracción de eventos: Consiste en identificar la aparición de eventos de un tipo particular.

El caso típico de extracción de información puede ejemplificarse con la tarea propuesta en la *Message Understanding Conferences* de 1991 (MUC-3). En esa ocasión, la tarea consistió en extraer, de forma automática, información específica desde reportes relacionados con terrorismo obtenidos de agencias de noticias de América Latina. Los sistemas participantes en esa conferencia tenían que llenar una plantilla con información como fecha, lugar y tipo del incidente, nombre del perpetrador, objetivo físico e instrumento usado para el ataque (Chinchor, Lewis, & Hirschman, 1993).

Según Hotho, Nürnberger, y Paab (2005), la extracción de información se realiza a través de una serie de procesos sobre los textos. Entre estos procesos están la tokenización (división del texto en unidades, generalmente palabras) y el etiquetado Part of Speech (asignación de la clase de palabra y categoría gramatical a la que pertenece cada palabra de un texto).

2.6. Clasificación

Basado en el trabajo de Gorunescu (2011), se puede decir que la clasificación es una técnica del aprendizaje automático que consiste en que dada una lista de categorías a la que puede pertenecer un elemento, este se relaciona a una de ellas según sus atributos. En otras palabras, los valores de sus atributos en conjunto definen la categoría a la que pertenece un elemento.

Para tal efecto, un conjunto de datos de entrenamiento ya clasificados manualmente es ingresado a un algoritmo para que éste identifique las características que hacen que un elemento (llamado ejemplo) pertenezca a una clase específica. La salida del algoritmo es un modelo de clasificación generado con los datos de entrenamiento que describe las condiciones que debe cumplir cada elemento para ser catalogado como un elemento de cierta clase. Posteriormente, el modelo de clasificación es probado con otro conjunto de datos también clasificados previamente de manera manual, llamados datos de prueba, para evaluar su desempeño.

Todo el procedimiento puede hacerse repetidamente, modificando el algoritmo, el volumen de datos de entrenamiento y de prueba, o la cantidad de atributos, para obtener un modelo eficiente. Finalmente es posible ingresar un conjunto de elementos nuevos, de los que no se conoce su clase, para que el modelo de clasificación la determine automáticamente; es decir, para que el modelo clasifique los nuevos ejemplos (predicción).

A continuación se enlistan algunos métodos de clasificación identificados por Gorunescu (2011) como de uso común. Cabe mencionar que cada uno de ellos integra diversos algoritmos para generar los modelos de clasificación.

a. Árboles de decisión.
b. Clasificadores Bayesianos.
c. Clasificadores basados en el vecino más cercano.
d. Métodos basados en reglas.
e. Máquinas de vector soporte.

2.7. Tecnologías del lenguaje en la administración del conocimiento

En el año 2003, Agustí Canals intuyó que surgirían nuevas tecnologías que influirían en la forma en que se ejecutan las tareas de la administración del conocimiento. Distinguió entre estas tecnologías las siguientes: el descubrimiento de conocimiento en BD o minería de datos, los sistemas pregunta-respuesta y los sistemas de búsqueda y recuperación de información (Canals, 2003).

Posteriormente Vorakulpipat y Rezgui (2008) reconocieron la importancia que ciertos desarrollos tecnológicos tienen como elementos clave en la creación de valor mediante la administración del conocimiento. Estos autores aseguraron que tecnologías como la web semántica, el procesamiento del lenguaje natural (PLN), las herramientas colaborativas y las tecnologías móviles eran importantes para el futuro de esta disciplina.

El mismo autor asegura que la inteligencia artificial será uno de los elemento que trazarán el futuro desarrollo de la administración del conocimiento, entre otras razones, porque permite la creación de nuevas herramientas que automatizan el proceso cognitivo humano. Un ejemplo de publicaciones científicas que confirman la intuición de los autores mencionados anteriormente son los trabajos de Diao, Zuo, y Liu (2009).

Por su parte, Maybury (2001) exhibe que el procesamiento del lenguaje natural, como parte de la inteligencia artificial, puede proveer herramientas de recuperación, extracción, resumen, traducción, presentación y generación de conocimiento. Esta autor muestra que los mayores avances se han logrado en la creación de herramientas avanzadas de búsqueda y sistemas de localización de expertos.

Existen varios ejemplos de la aplicación de las tecnologías del lenguaje en la administración del conocimiento. Por ejemplo, Maybury (2001) presenta un buscador de expertos, desarrollado para la *National Aeronautics and Space Administration* (**NASA**) por la universidad del estado de Florida, que, de manera automática, extrae y correlaciona información para determinar los conocimientos especializados que posee un grupo de investigadores.

Este desarrollo realiza, en primer lugar, una extracción de entidades nombradas (tarea que forma parte de la extracción de información) desde las publicaciones creadas por los investigadores y desde aquellas que hacen referencia a su persona. Luego genera automáticamente el perfil de cada investigador y, finalmente, lo presenta en una lista ordenada por el perfil que más coincide con un término de búsqueda que ingresa algún usuario.

Otros ejemplos de desarrollos que integran múltiples tecnologías del lenguaje pueden encontrarse en trabajos como los de Moldovan (2001), Feldman, y otros, (1998), Liu, Liu, y Yang, (2008), Mansingh, Osei-Bryson, y Reichgelt (2009), Tedmori y Jackson (2012), por mencionar algunos.

2.8. Representación de conocimiento con ontologías

Badia Cardús (2003) asegura que la representación del conocimiento es una tarea fundamental en el procesamiento de lenguaje natural debido a que influye en los resultados del procesamiento del texto, y a la necesidad de representar el razonamiento y la inferencia para permitir representar el conocimiento del mundo, esto es, el conocimiento relativo al contenido de los textos que se procesan.

Para lograr esto último, se han aprovechado diversos elementos, entre los que se encuentran la lógica, las representaciones procedurales (ampliamente utilizadas en los sistemas expertos) y las redes semánticas. Las ontologías son un ejemplo de técnicas de representación de conocimiento de tipo red semántica, en ella es posible representar un conjunto de conceptos y las relaciones entre estos, en una estructura jerárquica, también llamada taxonómica (Kodratoff, 1999).

La palabra *ontología* es un término filosófico (Maedche, 2003). Sin embargo, en el área de la inteligencia artificial se puede decir que una ontología es una especificación formal y clara (estructura con reglas) de un modelo o sistema de conceptos que es generalmente aprobada por la mayoría de las personas (Borst & Akkermans, 1997).

La estructura de las ontologías hace visible, entre otras cosas, las relaciones de hiponimia y meronimia entre conceptos, lo que facilita las tareas de inferencia (Badia Cardús, 2003). La siguiente imagen muestra ejemplos de estas relaciones.

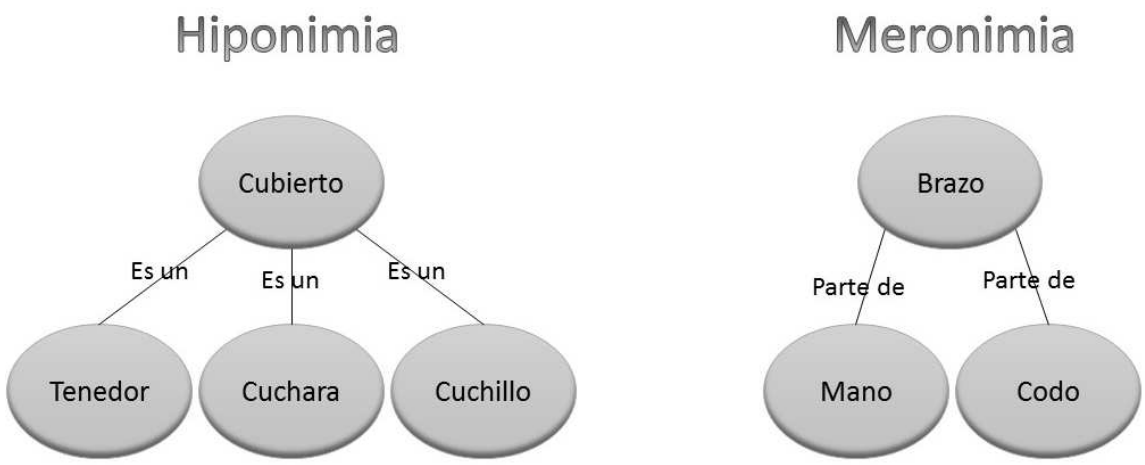


Figura 2-2 Relaciones de hiponimia y meronimia en ontologías.

Fuente: Elaboración propia

En los ejemplos anteriores, la información que caracteriza a los conceptos superiores de la estructura también es válida para los elementos inferiores con los que se tiene una relación, de manera que es posible codificar mucha información sin necesidad de repetirla en cada elemento. A esta característica se le llama herencia. Cardús (2003) afirma que la estructura de las ontologías permite una representación del conocimiento sencilla, clara e intuitiva.

Las ontologías generalmente son construidas manualmente. Sin embargo, existen trabajos que aplican las tecnologías del lenguaje para automatizar su construcción (Vorakulpipat & Rezgui, 2008). Ejemplos de este tipo de trabajos pueden encontrarse en la propuesta de Kodratoff (1999). En este puede observarse la aplicación de dichas tecnologías para encontrar automáticamente las relaciones taxonómicas y semánticas entre conceptos.

En la disciplina de la administración del conocimiento, las ontologías son cada vez más usadas ya que, como lo aseguran Liu, Liu, y Yang (2008), estas permiten que los documentos que registran el conocimiento puedan ser estructurados adecuadamente y semánticamente definidos, lo que facilita la recuperación de información pues posibilita la entrega de documentos con información relacionada a algún concepto buscado. Lo anterior

mejora la recuperación de información realizada con motores de búsqueda basados en palabras clave, ya que, como el mismo autor lo asegura, es imposible obtener toda la información necesaria y relacionada, utilizando solo palabras.

CAPÍTULO 3. Implementación de la administración del conocimiento un grupo de investigación

3.1. Antecedentes

Como ya se ha mencionado, en este trabajo de investigación se pretende aplicar tecnologías del lenguaje en la administración del conocimiento, específicamente en el repositorio de conocimiento que se produce tras su implementación. Para tal efecto, primero se realizó dicha implementación en una organización.

Debido a que otro de los objetivos de este trabajo es la observación de la implementación de la administración del conocimiento en grupos de investigación, se empleó como caso de estudio al Grupo de Ingeniería Lingüística (GIL) de la Universidad Nacional Autónoma de México (UNAM). Este grupo de investigación ha propuesto un proyecto para crear un Centro Nacional de Conocimiento, Información y Tecnologías del lenguaje, de manera que la implementación de la administración del conocimiento en esta entidad parece oportuna y apropiada.

En las siguientes secciones se presentará a dicho grupo de investigación y se describirá su proceso de implementación de la administración del conocimiento.

3.2. Caso de estudio: El Grupo de Ingeniería Lingüística de la UNAM

3.2.1. El Grupo de ingeniería lingüística de la UNAM

El Grupo de Ingeniería Lingüística de la UNAM (GIL) es una entidad dedicada a la investigación en torno a la lingüística computacional, el procesamiento de lenguaje natural y desarrollo de sistemas informáticos que puedan reconocer, comprender, interpretar y generar lenguaje humano, es decir, la ingeniería lingüística.

Este grupo surge en el año de 1999, teniendo como fundador al Dr. Gerardo Sierra Martínez, y gracias al apoyo del Instituto de Ingeniería y al Consejo Nacional de Ciencia y

Tecnología (CONACyT), este grupo realiza desde entonces proyectos relacionados al procesamiento del lenguaje natural (Sierra Martínez, 2009).

Desde esa fecha y hasta el día de hoy, el GIL, a través de su equipo multidisciplinario que integra a académicos y alumnos de computación, lingüística, informática, ingeniería, investigación de operaciones y bibliotecología, ha producido un número considerable de publicaciones, entre las que se encuentran tesis, artículos, reportes técnicos, etc.

Sierra Martínez (2009) señala que el GIL trabaja frecuentemente en colaboración con otras organizaciones dedicadas a la investigación, a continuación se enlistan algunas de ellas.

Dentro de la propia UNAM:

- Centro de Ciencias Aplicadas y Desarrollo tecnológico
- Instituto de Investigaciones Filosóficas
- Facultad de Filosofía y Letras
- Instituto de Matemáticas Aplicadas y en Sistemas

A nivel nacional:

- Instituto Politécnico Nacional (IPN)
- Benemérita Universidad Autónoma de Puebla
- El Colegio de México A. C.

A nivel internacional:

- Universidad de Manchester
- Universidad Pompeu Fabra
- Universidad de Aviñón
- Universidad Católica de Chile

Cabe mencionar que el GIL ha diseñado y presentado una propuesta para la creación de un Centro Nacional de Conocimiento, Información y Tecnologías del Lenguaje. En los trabajos de Sierra (2006) se señala que este proyecto responde a la necesidad de concretar un proyecto nacional para facilitar una gestión de la información que contribuya al descubrimiento, generación y administración del conocimiento, mediante el uso de tecnologías del lenguaje.

Lo anterior hace suponer que la implementación de la administración del conocimiento y la consecuente aplicación de las tecnologías del lenguaje en esta, es oportuna y acorde con las necesidades e intereses del grupo de investigación. A continuación se muestra la liga URL al sitio web del GIL, en ella se puede consultar más información referente a este grupo de investigación: http://www.iling.unam.mx.

3.3. Proceso de implementación

Actualmente existen diferentes modelos de implementación de la administración del conocimiento en organizaciones, también llamados marcos de trabajo o Frameworks, que proveen una guía para insertar el conjunto de actividades necesarias para la administración del conocimiento. Sin embargo, muchos de ellos no están disponibles de manera gratuita y otros están diseñados para utilizarse en organizaciones de gran tamaño o en organizaciones de tipo comercial, de manera que eran poco apropiadas para los fines del presente trabajo.

Para la implementación de la administración del conocimiento en el GIL se consultó como referencia al marco de trabajo de la APO, conjuntamente con la amplia literatura que constituye el estado del arte en la materia. Esto, debido a que existe acceso libre a su documentación oficial y a que existe literatura que comprueba su éxito en organizaciones de diversa índole, como gubernamentales y educativas, diversos giros económicos y diferentes tamaños.

La liga URL que se muestra a continuación corresponde al sitio web de la APO, desde el cual puede consultarse y obtenerse la documentación oficial completa del marco de trabajo mencionado, así como manuales, artículos, material para capacitación, publicaciones y otros elementos que facilitan la comprensión del framework y su utilización: http://www.apo-tokyo.org/publications/.

A continuación se muestra el diagrama que se encuentra en la literatura del framework mencionado y que resume y exhibe gráficamente dicho marco de trabajo.

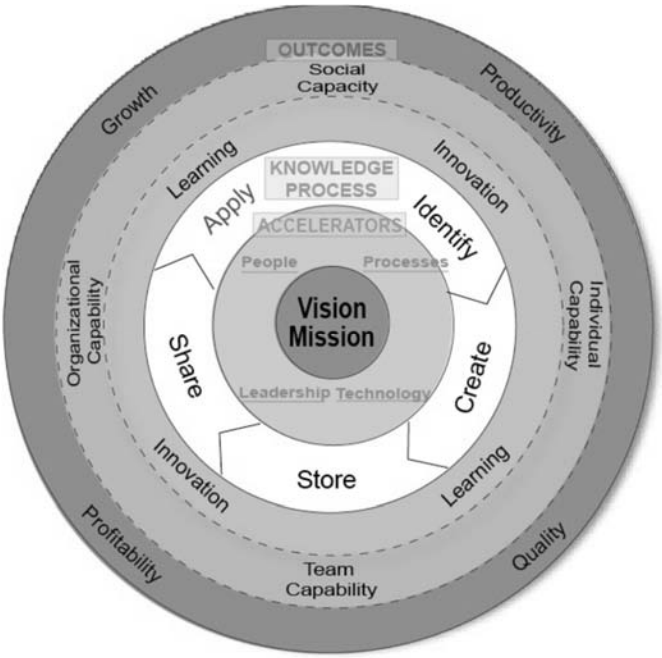


Figura 3-1 Marco de trabajo de la APO.

Fuente: Asian Productivity Organization (2009)

3.3.1. Creación del Plan de proyecto

El proceso de implementación de la administración del conocimiento en el GIL comenzó por la creación de un plan de proyecto que sirvió para especificar y documentar las características del proyecto con relación a los objetivos, alcances, recursos, actividades y manejo de riesgos. Esto facilitó el control y evaluación del progreso de este.

El plan de proyecto creado contiene, entre otras cosas, los siguientes elementos:

- Misión del proyecto
- Visión del proyecto
- Objetivos y metas a alcanzar
- Alcances y restricciones
- Beneficios esperados
- Análisis costo - beneficio
- Participantes en el proyecto y sus responsabilidades (Matriz RACI)

- Plan de actividades a realizar y sus tiempos de entrega
- Recursos a utilizar
- Manejo de riesgos (Controles)

Cada elemento de la lista anterior puede consultarse a detalle en el Anexo 1.

3.3.2. Carta aceptación del proyecto

El siguiente paso en el proceso de implementación de la administración del conocimiento en el GIL fue la firma de la carta aceptación del proyecto. Este documento avala el deseo de los directivos del GIL para realizar dicha implementación y compromete a los involucrados en el proyecto a colaborar comprometidamente en las actividades que requiera el proyecto. También sirve como evidencia de que los directivos del GIL han analizado el plan de proyecto presentado con anterioridad y otorgan su visto bueno.

En el Anexo 2 puede consultarse la carta aceptación del proyecto (por razones de seguridad se omite la imagen del documento original para que no aparezcan las firmas). Cabe mencionar que los marcos de trabajo para la implementación de la administración del conocimiento no contemplan esta actividad. Sin embargo, es una buena práctica de la administración de proyectos. Por esta razón se decidió realizarla.

3.3.3. Evaluación inicial

Una vez que se establecieron, formalizaron y aprobaron las características del proyecto (en los dos pasos anteriores), se realizó una evaluación de la situación actual de la organización con relación a los procesos de la administración del conocimiento. Esto se hizo considerando que existen trabajos de investigación, como los del Centro de sistemas de conocimiento del Tecnológico de Monterrey (2001), que muestran que es probable que algunas organizaciones realicen muchas de las actividades de la administración del conocimiento sin identificarlas formalmente con ese nombre.

Para hacer la evaluación inicial se aprovechó una herramienta publicada en la Asian Productivity Organization (2010) llamada APO KM Assessment Tool. Esta herramienta

permite conocer si la administración del conocimiento ya se practica en la organización evaluada y, en caso de ser así, en qué grado se aplica. También, facilita saber si la organización tiene las condiciones necesarias para implementar un proceso sistemático de administración del conocimiento. Además, en caso de que la organización ya siga un proceso sistemático de administración del conocimiento, la herramienta hace visible qué elementos pueden fortalecerse para mejorarlo.

La herramienta está conformada por una batería de 42 reactivos que evalúa siete elementos o categorías para determinar el grado en que estas apoyan a la administración del conocimiento de la organización, estas son:

1) Liderazgo
2) Procesos de la organización
3) Personas (Integrantes)
4) Tecnología
5) Procesos de Conocimiento
6) Aprendizaje e Innovación
7) Administración de los resultados de la administración del conocimiento

Los 42 reactivos (siete por cada categoría) son frases redactadas en modo afirmativo que indican la realización de una actividad en la organización. Cada integrante de la organización al que se le aplique la herramienta debe colocar un número, al frente de cada reactivo, que indique el grado de intensidad con que se realice cada actividad, según la escala siguiente.

Descripción	Escala de calificación
Haciéndolo muy bien	5
Haciéndolo bien	4
Haciéndolo adecuadamente	3
Haciéndolo limitadamente	2
Haciéndolo muy limitadamente o no se hace nada	1

Cabe mencionar que la herramienta APO KM Assessment Tool está redactada en inglés y no existe una traducción oficial en español. Se decidió no aplicar la herramienta en el idioma original (inglés) para que pudiese ser contestada por cualquier integrante del GIL sin necesidad de utilizar traductores automáticos o diccionarios inglés – español y, así, evitar

diferentes interpretaciones de los mismos conceptos. Por tanto, se realizó una traducción teniendo cuidado que esta se apegara lo más posible al texto original.

De igual manera, es necesario señalar que la herramienta APO KM Assessment Tool usa palabras relacionadas en mayor medida con organizaciones de giro comercial (por ejemplo, la palabras *gerente, clientes, rentabilidad, procesos del negocio*). Sin embargo, se decidió utilizar esta herramienta sin mayores adaptaciones ya que uno de los objetivos de este trabajo de investigación es observar el grado de transparencia con que se puede implementar la administración del conocimiento en grupos de investigación y las adaptaciones que deben hacerse. En consecuencia, se redactaron instrucciones de llenado básicas y se permitió que el aplicador de la herramienta aclarara dudas relacionadas con este tipo de frases o palabras. La herramienta completa, utilizada en la evaluación inicial, puede consultarse en el Anexo 3.

La documentación oficial de la herramienta indica que esta debe ser aplicada a diversas áreas de la organización, para que al final se elabore un diagrama de los resultados por área. También sugiere que las personas a las que se les aplique la evaluación tengan más de seis meses laborando en la organización con la finalidad de que estén familiarizados con sus procesos (Asian Productivity Organization, 2009). Sin embargo, se omitieron ambas indicaciones, debido a que, en el GIL, la población con estas características es mínima. En cambio, se aplicó el cuestionario a toda la población con más de 3 meses de antigüedad y los resultados fueron evaluados y analizados como una sola área global. El número total de encuestados fue nueve.

Los puntajes obtenidos en las encuestas, así como sus promedios, pueden observarse en la siguiente tabla.

Encuestado	Subtotal por Categoría						
	1	2	3	4	5	6	7
1	11	18	13	27	17	25	18
2	20	20	19	19	17	24	17
3	18	16	16	18	10	17	7
4	13	9	17	16	11	24	6
5	12	17	19	18	11	30	15
6	10	15	11	21	13	21	15
7	15	11	10	26	8	15	13
8	6	12	8	17	7	12	6
9	13	8	14	11	12	22	6
Promedio por categoría	13.11	14	14.11	19.22	11.77	21.11	11.44
Suma de promedios	104.78						

Tabla 3-1 Puntajes obtenidos en la evaluación inicial

Fuente: Elaboración propia

La siguiente gráfica de radar presenta los datos anteriores de manera resumida.

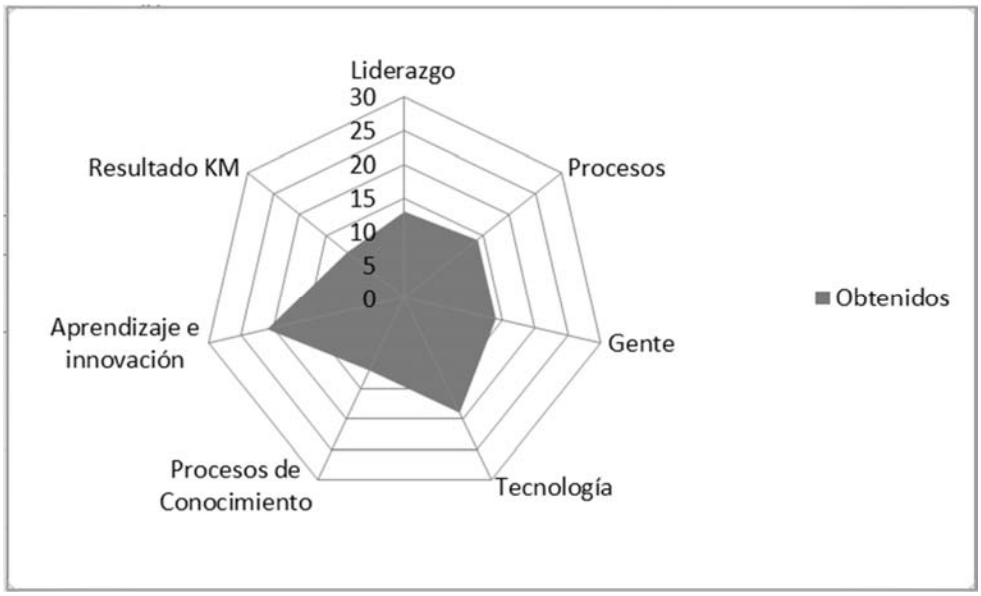


Figura 3-2 Gráfica de los puntajes obtenidos en la evaluación inicial

Fuente: Elaboración propia

Los datos de la tabla y, consecuentemente, los de la gráfica, muestran que la categoría más desarrollada es la de aprendizaje e innovación, con un valor de 21/30. Esto quiere decir que en el GIL se da una notoria importancia a estos elementos y se realizan cotidianamente acciones que promueven la innovación y el aprendizaje. Lo anterior tiene bastante sentido,

pues la organización analizada es un grupo de investigación y, por tanto, es natural que estos elementos tengan tal relevancia e impulso. En esta categoría se encontraron las siguientes fortalezas y oportunidades de mejora.

Fortalezas:

- Los equipos multidisciplinarios se organizan para hacer frente a los problemas que atraviesan las diferentes unidades de la organización.
- Se estimula a los colaboradores a trabajar en conjunto y compartiendo información.
- La organización comunica y refuerza continuamente los valores de aprendizaje e innovación.

Oportunidades de mejora:

- La organización considera tomar riesgos o cometer errores como oportunidades de aprendizaje, siempre y cuando no ocurran de manera repetida.

El elemento evaluado como el segundo más alto es la tecnología, con un valor de 19/30. Esto indica que en la organización existe una infraestructura tecnológica elemental para apoyar la creación, almacenamiento, transmisión y aplicación del conocimiento de la organización. En este rubro las fortalezas y oportunidades de mejora más relevantes encontradas son las siguientes.

Fortalezas:

- Todos los integrantes del GIL-UNAM tienen asignada una computadora con acceso a internet, intranet y servicio de correo electrónico.
- La dirección está dispuesta a probar nuevas herramientas y métodos que faciliten el trabajo y optimicen los resultados

Oportunidades de mejora:

- Existe una intranet pero no es utilizada como fuente principal de comunicación de toda la organización para apoyar la transferencia del conocimiento o el intercambio de información.
- La infraestructura de TI (Software y hardware) no permiten compartir eficientemente el conocimiento de la organización
- No existe un repositorio central que sirva como inventario de conocimiento de la organización.

Lo anterior indica que los procesos de aprendizaje e innovación y la tecnología, existentes en la organización, son los elementos que, al momento de la evaluación, aportan en mayor grado a la administración del conocimiento del GIL. Esto, a pesar de que este grupo de investigación no lo perciba formalmente con ese nombre.

Los datos resultantes también muestran que todas las categorías son susceptibles de mejorar. De manera que la implementación formal de la administración del conocimiento puede abordar la mejora de cualquiera de las categorías. Por lo tanto, la restricción en la amplitud de acción depende únicamente de los intereses, alcances y restricciones del proyecto de implementación.

En el caso particular, se decidió abordar con mayor énfasis, pero no de manera exclusiva, el desarrollo de la categoría de tecnología. Esto se debe a que el presente trabajo de investigación tiene la finalidad de aplicar tecnologías del lenguaje en el repositorio o portal (inexistente hasta el momento de la evaluación inicial) producido por la implementación de la administración del conocimiento. Las otras categorías también fueron abordadas, aunque en menor grado. Esto se hizo considerando la literatura existente que demuestra que la tecnología es un elemento importante en la administración del conocimiento, pero no lo es todo. Esta afirmación fue abordada con más detalle en el CAPÍTULO 1.

Cabe mencionar que la herramienta APO KM Assessment Tool proporciona más información. Sin embargo la mostrada con anterioridad fue la que se usó en mayor grado como referencia para observar el panorama general de la ejecución de procesos de la administración del conocimiento en el GIL, previo a su implementación formal.

3.3.4. Plan de cambio cultural

En los trabajos de Linkage (2000) se advierte que la principal dificultad en la implementación de la administración del conocimiento en organizaciones, según la experiencia de las organizaciones que ya hicieron la implementación, es el cambio cultural. Por esta razón, se diseñó un plan de cambio cultural y se incluyó en el plan de proyecto de implementación de administración del conocimiento en el GIL.

El plan de cambio cultural se diseñó tomando como referencia los trabajos de Kotter (1996). Este autor propone un conjunto de ocho pasos para minimizar el rechazo al cambio. A grandes rasgos, el plan estableció las siguientes actividades.

- Sostener diálogos con los integrantes del GIL para comunicar el proyecto y mostrar los cambios que se pretenden realizar.
- Asegurarse que los directivos del GIL apoyaran comprometidamente el proyecto de implementación de la administración del conocimiento.
- Compartir la visión del proyecto (documentada en el plan de proyecto) con todos los integrantes del GIL.
- Dejar abierta la posibilidad de recibir retroalimentación y exposición de dudas.
- Exhibir los logros obtenidos, conforme las metas del proyecto se vayan alcanzando.

3.3.5. Revisión de Misión, Visión, objetivos y procesos de la organización

La literatura existente sobre administración del conocimiento y la mayoría de sus frameworks advierten que la administración del conocimiento debe estar cuidadosamente alineada a la misión y visión de la organización.

Al momento de revisar la misión y visión del GIL, sus directivos percibieron que la redacción de estos no reflejaba con precisión la visión y misión real de la organización, de manera que decidieron redefinirlos. Al finalizar este proceso quedaron especificados de la siguiente manera.

- Misión: Generar conocimiento básico y aplicado en el área de la ingeniería lingüística.
- Visión: Ser el grupo líder en investigación y desarrollo tecnológico del área de la ingeniería lingüística.

En consecuencia, se establecieron nuevamente los objetivos de la organización para que estos quedaran alineados a la reciente misión y visón. Los objetivos de la organización, después de este proceso, pueden consultarse en el anexo 4.

Posteriormente se modificaron los procesos para asegurarse que estos apoyaran al alcance de los objetivos planteados y se documentaron. Esta documentación forma parte del conocimiento explícito del GIL, que posteriormente fue almacenado en un repositorio de conocimiento para garantizar que este conocimiento quedara disponible para quien necesite tenerlo. La creación del repositorio de conocimiento y el almacenamiento del conocimiento explícito en este serán descritos en una sección posterior.

En el transcurso de la modificación y documentación de los procesos de la organización se pudieron observar los cuatro tipos de creación de conocimiento expuestos en el trabajo de Nonaka (1994). El intercambio de ideas entre las personas que redefinieron los procesos corresponde a la socialización, la documentación de estos procesos concierne a la externalización, la revisión de diferentes documentos existentes para elaborar uno nuevo pertenece a la combinación, y la aceptación e integración del proceso rediseñado y plasmado en un documento representa la internalización.

3.3.6. Implementación de herramientas de la administración del conocimiento no basadas en TI

En el plan de proyecto se indica que una de las actividades a realizar es la implementación de herramientas de la administración del conocimiento no basadas en TI. Sin embargo, el GIL, sin saberlo, ya aprovechaba varias de ellas. Por ejemplo, los lunes de cada semana, el GIL tiene un espacio destinado para que cualquiera de sus integrantes pueda compartir algún tema de interés, presentar sus avances, inquietudes, propuestas, etc. y recibir retroalimentación. Este espacio, que el GIL llama *seminario*, es equivalente a las

herramientas de la administración del conocimiento llamadas Asistencia de pares, espacios físicos colaborativos físicos, lluvia de ideas y mentoría.

Otro ejemplo es que los integrantes del GIL se organizaron para tomar café a las 5:00 p.m., en ese momento aprovechan para platicar sus inquietudes, compartir experiencias, debatir ideas y solicitar ayuda a sus compañeros (conocidos en la administración del conocimiento como *pares*). Este espacio es equivalente a la herramienta conocida como cafés del conocimiento.

Después de observar este hecho, se deliberó que no era necesario implantar nuevas herramientas, simplemente se formalizaron como parte de la administración del conocimiento del GIL al presentarlas a sus integrantes con el nombre de herramientas de la administración del conocimiento.

3.3.7. Creación del repositorio de conocimiento

Una de las oportunidades de mejora que se encontraron en la evaluación inicial es que en el GIL no existe un repositorio central que sirva como inventario de conocimiento de la organización; únicamente cuenta con una intranet, pero esta contiene poca información y gran parte de ella no está actualizada. Además, esta intranet no es utilizada como fuente principal de comunicación para apoyar la transferencia del conocimiento.

Por su parte, mucha de la literatura existente sobre administración del conocimiento sugiere la creación de un repositorio de conocimiento en el que puedan agruparse las herramientas basadas en TI de la administración del conocimiento que se implementen para posibilitar o facilitar la realización de las actividades de identificar, crear, almacenar, compartir y aplicar el conocimiento de la organización.

Por estas razones, el siguiente paso en el proceso de implementación de la administración del conocimiento en el GIL fue la construcción de un repositorio de conocimiento. Para tal efecto, se decidió utilizar el sistema gestor de contenidos Joomla considerando lo siguiente:

- Es un software gratuito

- Existe documentación oficial muy completa
- Es un sistema web. Esto permite el repositorio de conocimiento del GIL sea accesible desde cualquier lugar con acceso a internet.
- El sitio oficial de Joomla cuenta con un foro donde se intercambia conocimiento y experiencias relacionadas con la instalación, configuración, mantenimiento y modificación del sistema
- En el sitio web del sistema existen desarrollos (software) oficiales, llamadas *extensiones,* que amplían las capacidades del sistema. Algunos de estos pueden ser aprovechados como herramientas de la administración del conocimiento del GIL. Por ejemplo, el foro, chat, gestores de documentos, sistemas de búsqueda avanzada, etc.
- La interfaz de usuario es relativamente fácil de modificar. Esto facilita que la interfaz del repositorio de conocimiento del GIL se ajuste a los lineamientos publicados por la Dirección General de Cómputo y de Tecnologías de la información (DGTIC) en el portal web http://recursosweb.unam.mx/recursos-web/lineamientos-unam/.
- Uno de los objetivos del presente trabajo de investigación es la aplicación de tecnologías del lenguaje en el repositorio de conocimiento creado. Esto obliga a que el código fuente sea accesible y modificable para hacer posible que el repositorio ejecute software previamente desarrollado, o que se pueda colocar directamente en él las líneas de código necesarias para procesar los textos del repositorio. El sistema Joomla está desarrollado utilizando PHP, HTML, CSS y XML, de manera que es posible realizar lo anteriormente descrito.

La documentación oficial del gestor de contenidos Joomla, así como el software base y los plug in se encuentran disponibles en la siguiente liga: http://www.joomla.org/.

Una vez que se tuvo instalado el gestor de contenidos Joomla, se modificó la interfaz de usuario para cumplir con los lineamientos de la Dirección General de Cómputo y de Tecnologías de Información y Computación (DGTIC) que es la entidad encargada de generar y comunicar estrategias basadas en Tecnologías de la Información y Comunicación y promover su adopción por parte de las demás entidades pertenecientes a la Universidad

Nacional Autónoma de México (entre las que se encuentra el GIL), para apoyar sus actividades (Secretaría General. Universidad Nacional Autónoma de México, 2012).

Luego, se creó la taxonomía del conocimiento para obtener la estructura adecuada para organizar el contenido del repositorio. Esto último se hizo utilizando la herramienta llamada *taxonomía del conocimiento*, propuesta por la Asian Productivity Organization (2009) y descrita en Asian Productivity Organization (2010). La estructura creada fue utilizada posteriormente para establecer la estructura del menú del repositorio de conocimiento.

La taxonomía creada puede observarse en el anexo 5. En ella puede observarse que existen elementos que se repiten en diferentes secciones. Esto es un resultado común de la taxonomía del conocimiento, la intención que se persigue es que el acceso al conocimiento se encuentre centralizado, es decir, que cada sección agrupe todo el conocimiento existente en la organización relacionada con ella, de manera que, si un elemento de conocimiento pertenece a dos secciones, desde ambas debe existir la forma de acceder a él. Cabe resaltar que el conocimiento no se duplica, sino solamente los accesos a él. Incluso, es factible (aunque no deseable) que parte del conocimiento se encuentre almacenado fuera del repositorio y que en este únicamente se encuentren los acceso a él.

Lo anterior facilita el control del acceso al conocimiento, pues es posible restringir el acceso de las personas a secciones específicas del repositorio de conocimiento, de manera que únicamente puedan consultar el conocimiento requerido para desarrollar sus actividades asignadas. Esto aporta al control del flujo del conocimiento, pues como se dijo en el capítulo I, es importante evitar fugas del conocimiento (parte del capital intelectual) de la organización.

Es necesario señalar que, a pesar de tener la posibilidad de hacer dichas restricciones, el GIL estableció como política de acceso al conocimiento del repositorio que todos los usuarios del este puedan acceder a todas las secciones. De manera que, todos los accesos del menú están activos para todas las cuentas de usuario activas del repositorio.

Para que un integrante del GIL tenga una cuenta activa es necesario que este llene un formulario de registro en el mismo repositorio y que el administrador de este active manualmente la cuenta.

Es posible que este proceso necesite ser mejorado para incrementar la seguridad del repositorio, sin embargo, eso queda fuera del alcance del presente trabajo de investigación, a manera de posible trabajo futuro.

3.3.8. Implementación de herramientas de la administración del conocimiento basadas en TI

La taxonomía del conocimiento que se construyó en el paso anterior sirvió para ubicar el conocimiento del GIL en las secciones adecuadas del repositorio de conocimiento. Sin embargo, en ese momento únicamente se tenía implementado el sistema base del repositorio, de manera que el siguiente paso fue integrar extensiones de Joomla que permitieran colocar y administrar el conocimiento de cada sección.

Para lograr esto, se agregaron gestores de documentos al repositorio. Estos se obtuvieron de la página oficial de Joomla (http://extensions.joomla.org/), así que no fue necesario codificar nuevo software, únicamente modificar algunas de sus características para adaptarlas a las necesidades del GIL. Estos gestores de contenido apoyan las actividades de almacenar, compartir y aplicar el conocimiento de la organización.

Posteriormente, se observó que era necesario implementar herramientas de comunicación y colaboración para facilitar y fomentar el intercambio de conocimiento entre los integrantes del GIL y, de paso, explicitar dicho conocimiento. Para tal efecto, se agregaron dos foros en el repositorio, mediante una extensión de Joomla, uno para discusiones sobre temas de investigación y otro para asuntos administrativos. La extensión aprovechada también se obtuvo de la página oficial de Joomla. El foro es una herramienta de la administración del conocimiento que apoya las actividades de crear, almacenar, compartir y aplicar el conocimiento.

Luego, se instaló una extensión que facilitó que varios integrantes (uno a la vez) modificaran un mismo documento colocado en línea. Esto facilitó el trabajo colaborativo para coordinar las actividades internas de difusión y de vinculación del GIL, así como para documentar proyectos.

Una vez que el repositorio de conocimiento del GIL tuvo instaladas las herramientas necesarias para permitir el almacenamiento de los documentos (ya existentes en la organización) que explicitan una parte del conocimiento de la organización, se solicitó a un grupo de sus integrantes (que desempeñaron el papel de *trabajadores del conocimiento*) que colocaran dichos documentos en las secciones correspondientes del repositorio. De esta manera, una gran parte del conocimiento explícito del GIL, existente hasta ese momento, quedó almacenada en su repositorio.

Cabe señalar que el sitio web de las extensiones oficiales para Joomla contiene muchas herramientas. La selección de las extensiones utilizadas en esta tesis se realizó atendiendo las necesidades particulares del GIL planteadas en las sesiones de trabajo con sus directivos. Además, se tomaron en cuenta las sugerencias, propuestas y retroalimentación de sus integrantes, brindadas en los espacios que se aprovecharon para comunicar el proyecto y sus avances.

3.3.9. Comunicación del proyecto y capacitación a los integrantes

Para mantener informados a los integrantes del GIL sobre el avance de la implementación de la administración del conocimiento, capacitarlos en el uso de las herramientas implementadas en el repositorio y atender sus inquietudes y propuestas, se aprovechó el seminario semanal, descrito en la sección 3.3.6.

A medida que se agregaban nuevas herramientas de la administración del conocimiento al repositorio, o cuando se añadía alguna funcionalidad, se presentaban en el seminario y se invitaba a que se utilizaran y se diera retroalimentación. En estos espacios también se explicó, de manera simple, teoría de la administración del conocimiento. La mayoría de las sesiones fueron acompañadas por los directivos del GIL.

CAPÍTULO 4. Explotación de repositorios de conocimiento mediante tecnologías del lenguaje

En los capítulos anteriores hemos visto cómo los repositorios producidos por la implementación de la administración del conocimiento facilitan las actividades de creación, almacenamiento y distribución del conocimiento de las organizaciones. Generalmente, el acceso a los documentos que contienen el conocimiento es a través de los motores de búsqueda integrados en el repositorio.

En este capítulo se muestra la posibilidad, oportunidad y viabilidad de aplicar las tecnologías del lenguaje para explotar los repositorios de conocimiento de una manera más eficaz. Para tal efecto, se exhibe el descubrimiento de nuevo conocimiento a través de la creación automática de conglomerados (clusters) de documentos, que proveen información sobre las tendencias de investigación del GIL. Estos documentos son extraídos del repositorio de conocimiento del GIL y procesados por tecnologías del lenguaje.

4.1. Corpus y herramienta para la creación de clusters

El repositorio de conocimiento del GIL almacena distintos tipos de documentos. En la experimentación realizada para la creación de clusters de documentos se utilizaron únicamente las tesis realizadas o dirigidas por los integrantes o exintegrantes del GIL. Estos documentos se encuentran en la sección *Publicaciones GIL* del repositorio de conocimiento de la organización (ver anexo 5).

En las tecnologías del lenguaje, al conjunto de documentos que son analizados se les llama corpus. El corpus de estudio de esta tesis está formado de 41 tesis en formato PDF, con una extensión promedio superior a las 100 páginas, con elementos textuales y gráficos. Cada documento contiene en promedio 50,000 palabras.

El corpus completo puede ser descargado desde el portal de conocimiento del GIL, siendo necesario un registro previo y la aprobación por parte del administrador del portal.

La herramienta de minería que se utilizó en algunos de los procesos de la creación de los clusters es WEKA, un software desarrollado por la universidad de Waikato en Nueva Zelanda y distribuido bajo el General Public Licence (GPL). Este software integra un

conjunto de algoritmos usados en la minería de datos y, por tanto, aplicables a la minería de textos. Estos algoritmos están organizados en 5 secciones según la tarea que desarrollan. Existe también una sexta sección dedicada a opciones de visualización gráfica.

Las 6 grandes secciones que integran este software son:

1. Preprocesamiento.

2. Clasificación.

3. Cluster.

4. Asociación.

5. Selección de atributos.

6. Visualización.

4.2. Método

Se observan tres grandes etapas en la metodología adoptada. Primeramente, un preprocesamiento de los documentos los transformó del formato original a uno más conveniente para ser manipulado por la herramienta de minería. Luego, la herramienta de minería fue aplicada con diferentes configuraciones y elementos para formar los clusters. Finalmente los clusters generados fueron revisados por un panel de expertos para interpretar los resultados. En las siguientes secciones se exhibirá con más detalle lo aquí descrito.

4.2.1. Preprocesamiento de los documentos

Para poder realizar el ejercicio de clustering sobre el corpus de tesis fue necesario aplicar un preprocesamiento que transformara los documentos (en formato PDF) a formato TXT para permitir su manipulación por los algoritmos de clustering. Esta conversión se realizó utilizando la herramienta PDFBOX. Esta herramienta y su documentación están disponibles en el siguiente enlace.

https://pdfbox.apache.org/

Posteriormente, a las tesis ya convertidas a formato TXT se les eliminaron los caracteres no alfanuméricos para reducir el número de elementos textuales existentes. Esta eliminación fue realizada por un script programado en el lenguaje de programación Python.

Finalmente, un segundo script programado en el mismo lenguaje generó un conjunto de matrices en las que cada renglón representa un documento del corpus y las columnas guardan las características de dichos documentos. En la minería de textos estas características corresponden generalmente al conjunto de palabras de todos los documentos (diccionario) y las celdas de las matrices corresponden a las frecuencias de aparición de cada palabra. Estas matrices fueron guardadas en archivos CVS para su posterior tratamiento por la herramienta de minería.

La caracterización de los atributos de los documentos (diccionario) se hizo de las cinco formas siguientes:

1. Frecuencias absolutas de aparición sin lista de paro (SW, por sus siglas en inglés de Stop Words). La frecuencia absoluta es el conteo de las veces que aparece cada palabra en un documento. Las listas de paro contienen un conjunto de palabras que son frecuentemente usadas en un idioma. La finalidad es que estas palabras no se contabilicen porque, de hacerlo, tendrían un valor muy alto pues aparecerían muchas veces en un documento.
2. Frecuencias absolutas de aparición con lista de paro.
3. Frecuencias relativas de aparición sin lista de paro. La frecuencia relativa es el número de veces que aparece cada palabra entre el número total de palabras contenidas en un documento
4. Frecuencias relativas de aparición con lista de paro.
5. Valor TF-IDF (Salton & Buckley, 1988). Esta medida considera el número de veces que aparece una palabra (llamada término) en un documento (TF) y la frecuencia con que esta aparece en los demás documentos de la colección (IDF). La intención es identificar la relevancia que tiene cada palabra. Para esto, se considera que una palabra con una frecuencia alta en un documento, y que también tiene una frecuencia alta en los demás documentos de la colección, es una palabra con poca relevancia. En cambio, una palabra que aparece frecuentemente en un documento, y escasamente en los otros

documentos de la colección, es una palabra que caracteriza al documento, es decir, es una palabra con alta relevancia.

La frecuencia de aparición de una palabra en los demás documentos de una colección debe expresarse con un valor que disminuya el valor asignado a la relevancia del término cuando dicha palabra aparece frecuentemente en dicha colección. Para tal efecto se usa el inverso, esto es, IDF (Manning, Raghavan, & Schütze, 2008).

El uso de estas medidas y pesos de palabras en la presente experimentación responde al hecho de que en el área de recuperación de información se ha observado la necesidad de usar medidas diferentes a las simples frecuencias de palabras, pues, como se pudo observar en el párrafo anterior, las frecuencias altas no siempre corresponden con palabras importantes o descriptivas de un documento.

Cabe mencionar que las matrices generadas tenían un número de columnas (dimensión) superior a 65,000. Esto es, el tamaño del diccionario de palabras.

La siguiente figura ilustra de manera gráfica lo aquí descrito.

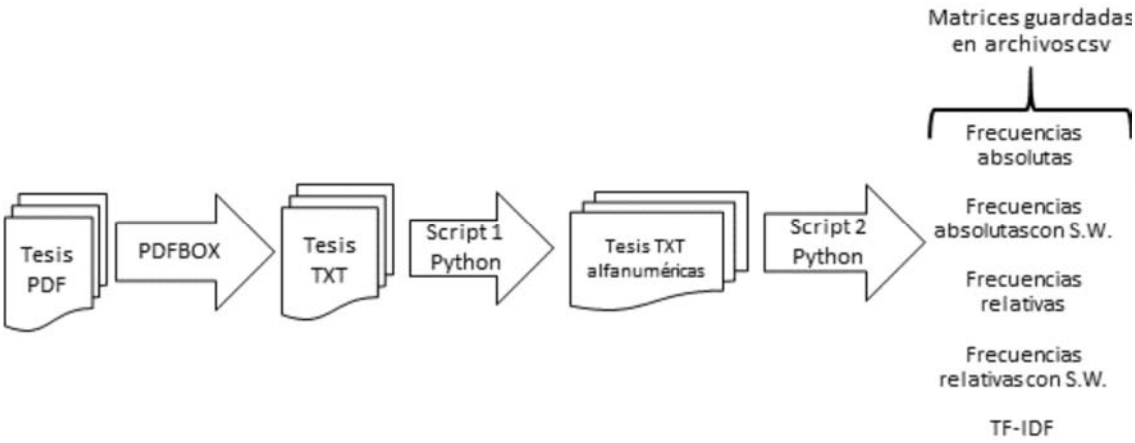


Figura 4-1 Preprocesamiento de los datos

Fuente: Elaboración propia

4.2.2. Experimentos

Para la experimentación se decidió usar dos algoritmos de clustering: el kMeans y el EM. El algoritmo kMeans fue elegido por su agilidad y buena reputación para tareas de clustering.

Por su parte, el algoritmo EM se eligió de manera estratégica como una forma de resolver el problema de no conocer de manera *a priori* el número de clusters que se debían formar.

En tareas de aprendizaje automático se recomienda dividir el conjunto de datos en dos partes. Una de ellas se utiliza para aprender el modelo (aprendizaje) y la otra para probar este modelo en datos distintos a los utilizados en el entrenamiento, es decir, aplicar el modelo. La idea detrás es que el modelo aprendido sea generalizable. Debido al número reducido de tesis se decidió utilizar el 50% de ellas como elementos de entrenamiento y el 50% restante para la creación de los clusters.

Se tomó cada una de las 5 matrices generadas en el preprocesamiento y se les aplicaron los dos algoritmos mencionados. En cada experimento se fijó un valor distinto en los parámetros de la configuración de los algoritmos. Finalmente se obtuvo un conjunto de 21 experimentos. Estos pueden consultarse en el anexo 6.

A continuación se expondrán los experimentos que se fueron efectuando y las salidas obtenidas. Estas salidas sirvieron para decidir el curso de la experimentación. Esta comenzó utilizando la matriz de frecuencias absolutas de aparición sin lista de paro.

Al inicio se utilizó el algoritmo kMeans y se observó que, sin importar el número de clusters que se le solicitara crear, el algoritmo asignaba casi la totalidad de los elementos a un solo cluster y a los demás únicamente les colocaba un elemento.

Esto hacía pensar que no existían grupos diferenciables en el conjunto de elementos. Lo mismo sucedía si se cambiaba el valor de la semilla o el porcentaje de entrenamiento. Sin embargo, al utilizar el algoritmo EM con los parámetros por default se encontró una distribución de los elementos en tres clusters principales. Esto facilitó concentrarse en experimentar de manera más específica, buscando encontrar un número de clusters cercanos a tres.

Los hechos anteriores pueden observarse en la siguiente tabla, que muestra una parte del conjunto de experimentos realizados.

Experimentos			
#	Algoritmo	Parámetros	Clusters obtenidos
2	Simple Kmeans	Número de clusters:3, Distancia: euclidiana, Semilla: Manual (6), Split: 50 % (entrenamiento) 50 % (Prueba)	Cluster 0: 0 u, Cluster 1: 20 u, Cluster 2: 1 u, Total: 21 unidades (50 %)
4	EM	Número de clusters: automático, Semilla: Manual (100), Split: 50 % (entrenamiento) 50 % (Prueba), MinStdDev:1.0 E-6	Cluster 0: 1 u, Cluster 2: 7 u, Cluster 3: 2 u, Cluster 4: 1 u, Cluster 5: 1 u, Cluster 7: 9 u, Total: 21 unidades (50 %)

Tabla 4-1 Experimentos con matriz de frecuencias absolutas sin lista de paro[1]

Fuente: Elaboración propia

Posteriormente, se experimentó con la matriz de frecuencias absolutas con lista de paro. Se observó que los resultados eran muy similares a los obtenidos con la matriz de frecuencias absolutas sin lista de paro. La siguiente tabla muestra un ejemplo de este fenómeno.

Experimentos			
#	Algoritmo	Parámetros	Clusters obtenidos
7.1	EM	Número de clusters: Automático, Semilla: Manual (100), Split: 50 % (entrenamiento) 50 % (Prueba), MinStdDev:1.0 E-6	Cluster 0: 1 u, Cluster 1: 7 u, Cluster 2: 2 u, Cluster 3: 1 u, Cluster 4: 1 u, Cluster 5: 9 u, Total: 21 unidades (50 %)

Tabla 4-2 Experimento con matriz de frecuencias absolutas con lista de paro

Fuente: Elaboración propia

[1] De manera predeterminada, el algoritmo EM oculta los clusters con cero elementos. Por esta razón en la gráfica no aparecen los clusters 1 y 6

A continuación se experimentó con la matriz de frecuencias relativas sin lista de paro y se encontró que el algoritmo kMeans seguía sin poder encontrar clusters en los elementos. Por su parte, el algoritmo EM encontraba más similares los elementos y los agrupaba en dos clusters principales cuando se le indicaba una mínima desviación estándar de 1.0 E-6 (valor por default del algoritmo).

En consecuencia, el siguiente experimento consistió en disminuir la desviación estándar mínima hasta 1.0 E-12. Esto permitió verificar que al obligar al algoritmo a diferenciar los elementos a un nivel de detalle mayor, se obtenían los mismos 3 clusters de los experimentos anteriores.

Este último experimento permitió observar que existía un elemento outlier[2] que formaba siempre un cluster con ese único elemento. Este elemento fue retirado para observar qué sucedía con la experimentación siguiente sin su presencia. La siguiente tabla muestra este hecho.

Experimentos			
#	Algoritmo	Parámetros	Clusters obtenidos
9	EM	Número de clusters: 3 Distancia: euclidiana Semilla: Manual (6) Split: 50 % (entrenamiento) 50 % (Prueba) MinStdDev:1.0 E-6	Cluster 0: 0 u Cluster 1: 18 u Cluster 2: 3 u Total: 21 unidades (50 %)
11	EM	Número de clusters: automático Distancia: euclidiana Semilla: Manual (6) Split: 50 % (entrenamiento) 50 % (Prueba) MinStdDev:1.0 E-12	Cluster 0: 10 u Cluster 1: 1 u Cluster 2: 7 u Cluster 3: 3 u Total: 21 unidades (50 %)

Tabla 4-3 Experimentos con matriz de frecuencias relativas sin lista de paro

Fuente: Elaboración propia

[2] Un outlier es un elemento que tiene características diferentes a los demás elementos de una colección. Para el caso específico de la presente experimentación, el elemento outlier corresponde a un documento que trata una temática notablemente distinta a la temática tratada por los demás documentos.

Una vez eliminado el outlier se utilizó la matriz de frecuencias relativas con lista de paro. Los resultados anteriores hacían intuir que los resultados no variarían con el uso de esta matriz con lista de paro. Esta intuición fue confirmada con el experimento número 14, que se expone en la siguiente tabla. Puede observarse que el resultado es el mismo, a excepción de que, al eliminar el elemento outlier, el número de clusters disminuyó en una unidad.

Experimentos			
#	Algoritmo	Parámetros	Clusters obtenidos
14	EM	Número de clusters: automático Semilla: Default (100) Split: 50 % (entrenamiento) 50 % (Prueba) MinStdDev:1.0 E-12	Cluster 0: 10 u Cluster 1: 7 u Cluster 2: 3 u Total: 20 unidades (50 %)

Tabla 4-4 Experimento con matriz de frecuencias relativas con lista de paro

Fuente: Elaboración propia

En seguida, se utilizó la matriz con los valores TF-IDF y se realizaron experimentos con los dos algoritmos, intuyendo que con esta matriz el algoritmo kMeans tendría más facilidad para diferenciar y agrupar los elementos en comparación con las matrices anteriores. Sin embargo, al ejecutar los experimentos, el algoritmo siguió sin poder diferenciarlos y entregó nuevamente un solo cluster. Este resultado no cambió al modificar las distancias, semillas y porcentaje de entrenamiento.

Para el caso del algoritmo EM, los resultados de los experimentos fueron consistentes con los anteriores, es decir, encontró los mismos tres clusters y éstos mantenían la misma proporción de elementos dentro de cada uno.

Lo anterior puede observarse en la siguiente tabla.

Experimentos			
#	Algoritmo	Parámetros	Clusters obtenidos
16	EM	Número de clusters: automático Semilla: Default (100) Split: 50 % (entrenamiento) 50 % (Prueba) MinStdDev:1.0 E-6	Cluster 0: 14 u Cluster 1: 4 u Cluster 2: 1 u Total: 19 unidades (50 %)
20	Kmeans	Número de clusters: 3 Distancia: Manhattan Semilla: Manual (10) Split: 50 % (entrenamiento) 50 % (Prueba)	Cluster 0: 19 u

Tabla 4-5 Experimentos con matriz TF-IDF

Fuente: Elaboración propia

Finalmente, se tenía la intuición de que el trabajo de los algoritmos podría optimizarse si se eliminaban las columnas con todos sus valores en cero de la matriz de valores TF-IDF. Esta intuición estaba basada en el hecho de que todos los elementos tenían muchos valores cero en sus vectores (matrices dispersas) y esto los hacía parecerse mucho. Si se eliminaban dichos valores, los datos restantes diferenciarían con más facilidad a los elementos.

Por tal motivo, se eliminaron las columnas que tenían todos sus valores en ceros y se realizó otro experimento. El resultado no cambió y el tiempo ocupado por el algoritmo disminuyó ligeramente. Por el contrario, el tiempo que ocupó el script que eliminó las columnas fue notablemente extenso.

La siguiente tabla muestra el resultado de este último experimento.

Experimentos			
#	Algoritmo	Parámetros	Clusters obtenidos
21	EM	Número de clusters: automático Semilla: Default (100) Split: 50 % (entrenamiento) 50 % (Prueba) MinStdDev:1.0 E-6	Cluster 0: 1 u Cluster 2: 3 u Cluster 3: 15 u Total: 19 unidades (50 %)

Tabla 4-6 Experimento con matriz TF-IDF sin columnas en ceros

Fuente: Elaboración propia

4.2.3. Resultados y evaluación

Los resultados de los experimentos mostraron consistentemente que existen tres grandes clusters que agrupan al conjunto de tesis. El paso siguiente, según la metodología establecida, era el análisis de los clusters por parte de un panel de expertos integrado por personas conocedoras de los temas y cercanas a la producción de las tesis analizadas.

Para tal efecto, se distinguieron los resultados más consistentes e identificaron los elementos de cada cluster encontrado. El producto fue un conjunto de 6 documentos; cada uno de ellos con tres clusters y sus respectivas tesis. Este conjunto fue entregado a los expertos para su interpretación. Cabe precisar que el análisis de los documentos por parte de los expertos se hizo de manera individual para comparar objetivamente las interpretaciones.

La siguiente tabla muestra el contenido de uno de los documentos. Cada columna muestra un cluster generado automáticamente. En cada cluster aparece, en primer lugar, el nombre del archivo digital de la tesis, enseguida aparece (entre paréntesis) el tipo de tesis y el título de la tesis.

El tipo de tesis se indica usando la siguiente nomenclatura:

- L= licenciatura
- M=Maestría
- D= Doctorado
- Ling= Lingüística

- IngComp=Ingeniería en computación
- Bibliotec= Boblioteconomía
- LyLh= Lengua y literatura hispánica
- CienComp= Ciencias de la computación

No. Experimento: 14		
0	1	2
Ecesarantonioaguilar_2008.txt (D-Ling) Análisis lingüístico de definiciones en contextos Definitorios Eirasemacruzdominguez_2011.txt (L-LyLh) EL SINTAGMA NOMINAL EN LA EXTRACCIÓN DE RELACIONES LÉXICO SEMÁNTICAS DE CONTEXTOS DEFINITORIOS EL CASO DE LA PREPOSICIÓN DE Ecarlosfranciscomendez_2009.txt (M-Ling) IDENTIFICACIÓN AUTOMÁTICA DE CATEGORÍAS GRAMATICALES EN ESPAÑOL DEL SIGLO XVI Emarinavladimirovnafomicheva_2012.txt (M-Ling) ANÁLISIS LINGÜÍSTICO DE LA TRADUCCIÓN AUTOMÁTICA PARA SU EVALUACIÓN Eximenagutierrezvazquez_2010.txt (L-IngComp) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO Ealexgarduñosandoval_2010.txt (L-IngComp) DESARROLLO DE UN CONSTRUCTOR AUTOMÁTICO DE BASES DE DATOS RELACIONALES A PARTIR DE ESQUEMAS XML Eadrianamirandanava_2006.txt (L-IngComp) GEOVARIANTES LÉXICAS DEL ESPAÑOL Ecarlosfranciscomendez_2013.txt (D-Ling) GENERACIÓN AUTOMÁTICA DE UNA GRAMÁTICA DE ESTADOS FINITOS PARA LA MORFOLOGÍA DEL ESPAÑOL Eitacruzsherling_2013.txt (L-IngComp) CONTEXTOS DEFINITORIOS EN LA EXTRACCIÓN DE TAXONOMÍAS EL CASO DE PLN Emaríajimenezvasques_2012.txt (L-IngComp) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO	Eluiscabreradiego_2011.txt (L-IngComp) TF IDF PARA LA OBTENCIÓN AUTOMÁTICA DE TÉRMINOS Y SU VALIDACIÓN MEDIANTE WIKIPEDIA Esistemasrecuperacioninformacion.txt (M-Ling) ESTRUCTURACIÓN SEMÁNTICO PRAGMÁTICA DEL LÉXICO EN DOMINIOS RESTRINGIDOS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN Ebrendacastrorolon.txt (L-LyLh) DETECCIÓN DE SIMILITUD TEXTUAL MEDIANTE CRITERIOS DE DISCURSO Y SEMÁNTICA Epavelsorionomorales_2011.txt (L-IngComp) CLASIFICACIO N DE OPINIONES MEDIANTE APRENDIZAJE DE MA QUINAS EL CASO DE RESEN AS SOBRE PELI CULAS Eariadnahernandezanguilo_2009.txt (L-LyLh) ANÁLISIS LINGÜÍSTICO DE DEFINICIONES ANALÍTICAS PARA LA BÚSQUEDA DE REGLAS QUE PERMITAN SU DELIMITACIÓN AUTOMÁTICA Eoctaviosanchezvelazquez_2009.txt (L-LyLh) LA FUNCIONALIDAD AL INTERIOR DE CONTEXTOS DEFINITORIOS CON DEFINICIONES ANALÍTICAS EL PATRÓN SINTÁCTICO PARA INFINITIVO Egabrielcastillohernandez_2012.txt (M-CienComp) ALGORITMO REVISADO PARA LA EXTRACCIÓN AUTOMÁTICA DE AGRUPAMIENTOS SEMÁNTICOS	Epérezpéreznuri.txt (M-Ling) PROBLEMAS DE LOS DICCIONARIOS ESPECIALIZADOS deltutado DE LIBRE COMERCIO DE AMÉRICA DEL NORTE CASO práctico SOBRE LA terminología REFERENTE A LA inversión EXTRANJERA Eazuryaparicioaguilar_2011.txt (L-Enfermería) ESTADO DEL CONOCIMIENTO SOBRE SEXUALIDAD HUMANA SALUD SEXUAL Y SALUD REPRODUCTIVA UN ESTUDIO BIBLIOMÉTRICO DE LA PRODUCTIVIDAD CIENTÍFICA EN ESPAÑOL DEL 2001 2010 Ejesusvaldesramos_2013.txt (D-Bibliotec) BASES LINGÜÍSTICAS E INFORMÁTICAS PARA LA ELABORACIÓN DE TESAUROS

Tabla 4-7 Ejemplo de documento entregado al equipo de expertos

Fuente: Elaboración propia

Los clusters formados por los experimentos más consistentes (con tres clusters formados) pueden observarse en el anexo 7.

El panel de expertos, compuesto por personal de investigación del GIL, con suficiente antigüedad y experiencia para opinar sobre la producción académica de la empresa, analizó los clusters y encontró que tres clasificaciones obtenidas tienen sentido (una de ellas es la mostrada en el cuadro anterior), aunque afirman que algunos de los elementos que aparecen en ciertos clusters no aparentan guardar relación con los demás. Los demás modelos de agrupamiento obtenidos no parecen mostrar un correcto agrupamiento, aunque existe la posibilidad de que el grupo de expertos no logre observar esa relación.

Los integrantes de dicho panel de expertos se percataron de que los clusters formados parecían agrupar documentos según la disciplina que estos abordan. El primer cluster agrupa tesis que fueron desarrolladas con una perspectiva mayoritariamente lingüística (que realizan un mayor análisis lingüístico). El segundo cluster conjunta tesis con una perspectiva mayoritaria de tecnologías del lenguaje (menor análisis lingüístico), principalmente tesis de ingeniería en computación

4.3. Conclusiones de la creación automática de clusters

Los clusters generados automáticamente sugieren que los trabajos realizados por la organización apuntan principalmente a dos áreas, la lingüística y las tecnologías del lenguaje. Además, se observa que esta última es la que ha tenido un mayor tratamiento.

Para este ejercicio de minería de textos, el resultado apreciablemente mejor se obtuvo con el uso de frecuencias relativas sin lista de paro ya que el uso de esta no cambió notablemente el resultado.

El uso de un procesamiento más completo como TF-IDF aumenta considerablemente el tiempo de procesamiento y no mostró una mejora visible en los resultados.

El algoritmo EM aparentemente no se ve afectado por columnas que tengan todos sus elementos en ceros, pues el resultado es exactamente el mismo.

El trabajo a futuro consiste en realizar experimentación usando algoritmos de clustering jerárquico para observar si existen subgrupos en los clusters encontrados que sugieran una agrupación más detallada de las tesis. También podría experimentarse usando "lemas de las palabras[3]" obtenidos mediante alguna herramienta de análisis lingüístico como el etiquetador FreeLing. Esto permitiría observar si la pérdida de los detalles afecta o no a la conglomeración de los elementos.

[3] Un lema es una palabra que representa un conjunto de variantes morfológicas. Para flexiones verbales se asigna el infinitivo, por ejemplo, para las variantes *corro, corren, corrieron, corramos, correrán* se asigna el lema *correr*. Para los sustantivos se asigna el masculino singular, por ejemplo, para *niño, niña, niños, niñas* se asigna *niño*. Existen herramientas de las tecnologías del lenguaje que realizan este proceso de forma automática. Por ejemplo el software FreeLing.

CAPÍTULO 5. Metodología para la organización semiautomática de nuevo conocimiento

5.1. Antecedentes

En secciones anteriores de este trabajo, se ha visto cómo la administración del conocimiento establece un conjunto de actividades generales. Estas actividades suelen apoyarse en herramientas que permiten un sostenido intercambio de conocimiento. Este intercambio facilita la creación de nuevo conocimiento y la actualización del mismo.

En esta dinámica de generación y actualización de conocimiento, es conveniente encontrar una forma de insertar el nuevo conocimiento en el lugar adecuado del repositorio para garantizar su disponibilidad y facilitar su acceso. La organización por áreas de conocimiento, visto en una sección anterior, permite alcanzar tal fin. Sin embargo, al ser un proceso continuo, demanda mucho esfuerzo y tiempo. En consecuencia, realizar este proceso de forma manual en ocasiones es impráctico, limitado y hasta imposible.

En este capítulo, se propone un proceso semiautomático para facilitar y aumentar la eficiencia de la organización del contenido de los repositorios de conocimiento. Para tal fin, se aprovechan las técnicas y modelos desarrollados en el área de las tecnologías del lenguaje. Estas servirán para procesar el conocimiento explicitado en las herramientas de la administración del conocimiento; herramientas que, a su vez, están integradas en los mismos repositorios.

Cabe mencionar que existen trabajos que aplican las tecnologías del lenguaje para crear herramientas avanzadas de búsqueda y sistemas de localización de expertos (Maybury, 2001). Sin embargo, se aprecia que existe una falta de herramientas que permitan la organización de repositorios de conocimiento de forma automática.

5.2. Organización semiautomática de nuevo conocimiento

Las siguientes secciones muestran la forma en que fue diseñada una metodología para organizar semiautomáticamente el conocimiento que se genera y explicita a través de algunas herramientas de la administración del conocimiento.

En este caso particular, se aprovechó el conocimiento creado, explicitado y almacenado recientemente en el repositorio de conocimiento del GIL, a través de la herramienta *foro*. En dicho repositorio existen dos foros, uno dedicado a cuestiones administrativas y otro destinado a temas de investigación. Para el diseño de esta metodología se utilizó únicamente el primero.

La idea central es que, cuando un integrante del GIL aporte nueva información a través del foro, el repositorio pueda reconocer automáticamente la sección a la que corresponde ese conocimiento, es decir, la sección donde este conocimiento debiera ser colocado, y mostrar la sugerencia al administrador del portal, o al usuario mismo. La hipótesis es que esto facilitaría y haría más eficiente la organización de los repositorios de conocimiento y, por tanto, mejoraría el acceso y entrega del conocimiento requerido por un usuario del mismo.

5.2.1. Modelado de las secciones del repositorio

Una forma de encontrar el lugar donde cierto conocimiento (explicitado en un documento) debe ser colocado, es aprovechando una técnica de aprendizaje automático llamada *clasificación.* Esta consiste en que se le entreguen a la computadora un abundante conjunto de datos, en nuestro caso comentarios del foro, previamente clasificados por un humano para que la computadora encuentre en ellos un modelo que sirva para clasificar los comentarios que se le entreguen en un futuro. Sin embargo, para nuestro caso de estudio esta técnica es inviable debido a que el número de comentarios en el foro del repositorio de conocimiento es limitado.

Otra alternativa consiste en observar y enlistar las entidades que caracterizan a cada elemento de la estructura del repositorio, a estas se les llamarán características. Luego, se toma cada comentario del foro y se identifica qué entidades (de las que se enlistaron anteriormente como características de los elementos del repositorio) aparecen en este. De esta manera, es posible comparar las entidades encontradas en el texto y las entidades que caracterizan a cada sección del repositorio. La premisa es que un comentario del foro tiene más probabilidad de pertenecer a una sección del repositorio mientras más características compartan.

Esta última opción permite trabajar con un número pequeño de textos, de manera que es la que se utiliza para la presente investigación. Por esta razón, el primer paso propuesto en esta metodología consiste en generar un modelo que describa las características de cada sección del repositorio de conocimiento. El modelado se puede hacer de varias maneras, utilizando diferentes técnicas de representación de conocimiento. Para nuestro caso de estudio se recurre al uso de ontologías.

A continuación se muestra una lista con la estructura del repositorio de conocimiento del GIL (después de la integración de las herramientas de la administración del conocimiento), que se diseñó aprovechando la taxonomía del conocimiento (ver sección 1.2.4.1).

Repositorio de conocimiento

- Capital humano
    - Personas
        - CV’s
    - Comunidades
        - Documentación
    - Redes de conocimiento
        - Documentación
- Investigación
    - Acervo bibliográfico
    - Bibtext
    - Corpus
    - Foro (temas de investigación)
    - Herramientas
    - Publicaciones GIL
    - Proyectos
- Formación de recursos humanos

    - Cursos
    - Publicaciones GIL
    - Acervo bibliográfico
    - Bibtext
- Difusión
    - Publicaciones GIL
    - Eventos (Para el público)
- Desarrollo de sistemas y tecnologías
    - Corpus
    - Herramientas
- Administración del laboratorio
    - Directorio
    - Asignación de recursos
    - Eventos internos
    - Expedientes
    - Foro para temas administrativos
    - Procedimientos
    - Galería
- Vinculación exterior
    - Eventos

La experimentación del presente trabajo se limita únicamente a clasificar el conocimiento del foro administrativo en los elementos de la sección *Administración del laboratorio* del repositorio de conocimiento. De esta manera, aunque en adelante se mencione el modelado del repositorio de conocimiento, se dará por entendido que únicamente se trabajará con la sección mencionada.

La siguiente imagen documenta el modelado de los elementos de la sección mencionada, a través de ontologías.

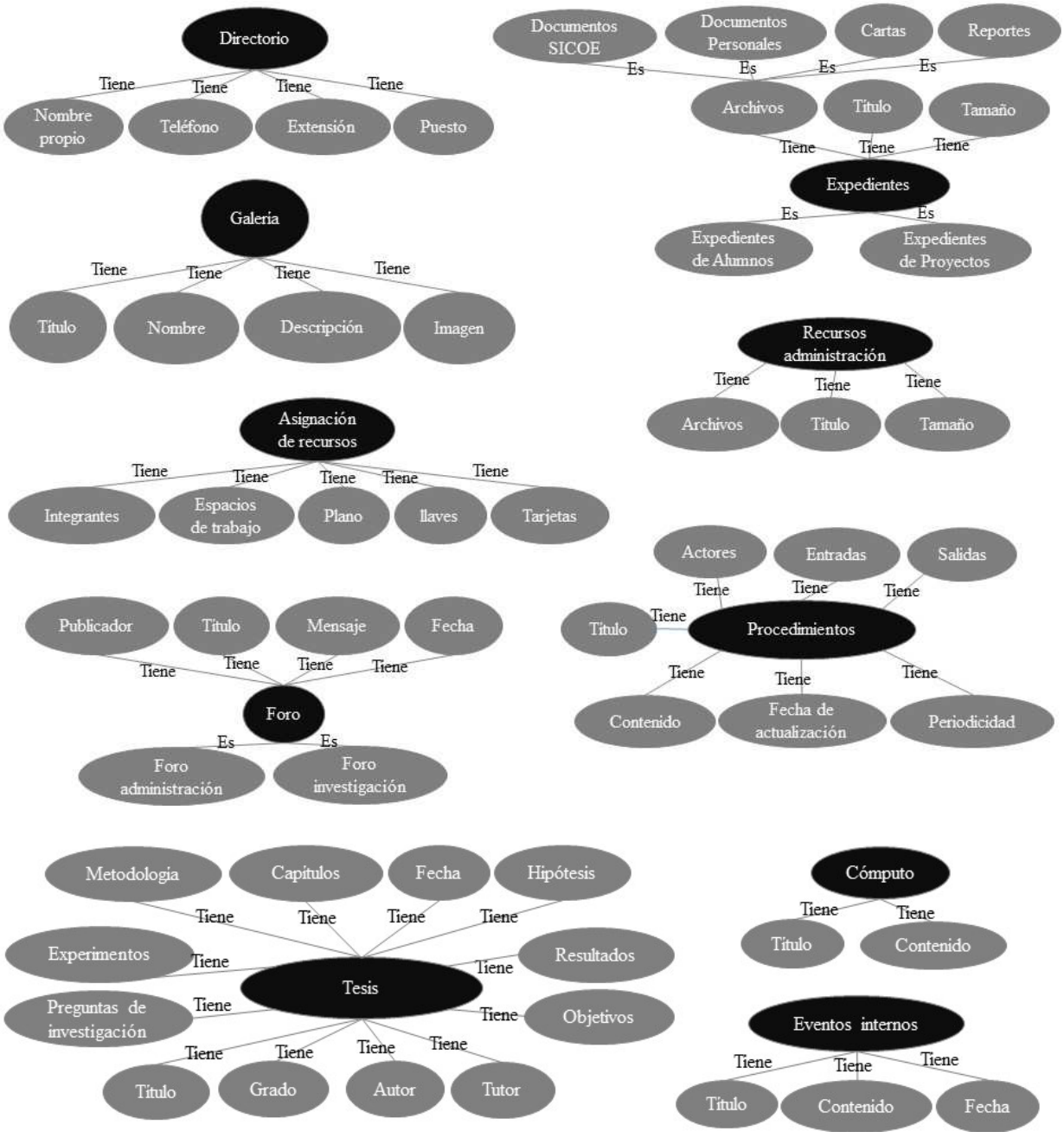


Figura 5-1 Modelado de los elementos de la sección Administración del laboratorio

Fuente: Elaboración propia

5.2.2. Uso de técnicas de las tecnologías del lenguaje

Se ha mencionado que el objetivo de las herramientas de la administración del conocimiento es proveer mecanismos que apoyen la realización de sus actividades generales para permitir un ágil y sostenido intercambio de conocimiento. Cada actividad se apoya de herramientas diferentes y no existe un estándar que defina cuáles herramientas deben ser utilizadas en cada caso. El uso de ellas depende de las necesidades que se deseen cubrir en cada implementación en específico. Sin embargo, es común encontrar el uso de blogs, chats y foros para apoyar las tareas de crear, almacenar y compartir el conocimiento de la organización.

Es necesario advertir que las herramientas mencionadas explicitan el conocimiento y lo almacenan en su mismo espacio, es decir en la herramienta misma. Por ejemplo, en un foro es fácil observar el intercambio del conocimiento (principalmente tácito) de muchos usuarios para solución de un problema. Dicha solución es un conocimiento que, en muchas ocasiones, no se encuentra plasmado en ningún otro lugar. Sin embargo, este conocimiento queda explicitado únicamente en el foro mismo, esto es, no se inserta este nuevo conocimiento en ningún otro espacio. Con esto, la forma de acceder al conocimiento generado queda definida por las tecnologías de búsqueda y acceso que integre la propia herramienta (normalmente a través de buscadores). Esto es poco conveniente, pues este conocimiento usualmente es nuevo o aporta elementos que actualizan al anterior. Por lo tanto, debería integrarse en los documentos que plasman el conocimiento de la organización para actualizarlos o expandirlos.

El problema de hacer esto de forma manual es que se requiere de una permanente lectura de los textos generados por la herramienta, la identificación del conocimiento relevante y su inserción en el documento adecuado. Este proceso demanda mucho esfuerzo y tiempo.

Las tecnologías del lenguaje han desarrollado métodos que permiten automatizar la tarea de lectura y extracción de información. La siguiente sección de este trabajo muestra la manera en que se usaron dichas tecnologías para procesar automáticamente los textos de las herramientas que explicitan el conocimiento de la organización.

5.2.3. Preprocesamiento de los textos

El conocimiento explicitado en los comentarios del foro del repositorio de conocimiento del GIL está plasmado en textos cortos con las siguientes medidas estadísticas:

Estadísticas por caracteres:

Extensión promedio (media): 219 caracteres
Desviación estándar de 202 caracteres
Rango: 38 a 1038 caracteres
Mediana: 169 caracteres
Moda: 94 caracteres

Estadísticas por palabras:

Extensión promedio: 36 palabras
Desviación estándar de 34 palabras
Rango: 4 a 179 palabras
Mediana: 28 palabras
Moda: 18 palabras

Los textos son guardados por el software del foro en la tabla *_discussions_messages* de la base de datos del gestor de contenidos (Joomla) que soporta al repositorio de conocimiento. Los textos se guardan específicamente dentro del campo *message* en formato de texto plano, de manera que se procesa información textual.

La siguiente tabla muestra el diseño de la tabla donde se guardan los comentarios del foro del repositorio de conocimiento:

Campo	Tipo	Nulo	Predeterminado
id	int(11)	No	
parent_id	int(11)	No	0
cat_id	int(11)	No	0
thread	int(11)	No	0
user_id	int(11)	No	0
account	varchar(50)	No	
name	varchar(100)	No	
email	varchar(100)	No	
ip	varchar(100)	No	
type	int(11)	No	1
subject	varchar(255)	No	
alias	varchar(255)	Sí	
message	text	No	
date	timestamp	Sí	CURRENT_TIMESTAMP
hits	int(11)	No	0
locked	tinyint(1)	No	0
published	tinyint(1)	No	0
counter_replies	int(11)	No	0
last_entry_date	timestamp	No	0000-00-00 00:00:00
last_entry_user_id	int(11)	No	
last_entry_msg_id	int(11)	No	
sticky	tinyint(1)	No	0
wfm	tinyint(1)	Sí	0
image1	varchar(255)	Sí	
image1_description	varchar(255)	Sí	
image2	varchar(255)	Sí	
image2_description	varchar(255)	Sí	
image3	varchar(255)	Sí	
image3_description	varchar(255)	Sí	
image4	varchar(255)	Sí	
image4_description	varchar(255)	Sí	
image5	varchar(255)	Sí	
image5_description	varchar(255)	Sí	
apikey_id	int(11)	No	0
latitude	float	Sí	NULL
longitude	float	Sí	NULL

Tabla 5-1. Estructura de la tabla _discussions_messages

Fuente: Elaboración propia

Para aplicar las tecnologías del lenguaje en el procesamiento del conocimiento inmerso en los comentarios del foro, es necesario realizar primeramente un conjunto de tareas de preprocesamiento de los textos. En primer lugar se extrajeron los comentarios de la base de datos a través de consultas SQL. Un ejemplo es el siguiente:

"SELECT message FROM _discussions_messages WHERE cat_id=6"

En seguida, se crearon tres archivos de texto por cada uno de los comentarios. El primero contiene el comentario original, tal cual se encuentra en la base de datos. El segundo guarda el texto alfanumérico que queda tras la eliminación de saltos de línea, dobles espacios y caracteres especiales como signos de puntuación, guiones, etc. El tercero contiene el texto del archivo anterior excluyendo, además, caracteres numéricos.

La razón de generar estas tres versiones de archivo del mismo comentario es que cada herramienta y método de las tecnologías del lenguaje tiene distinto resultado al trabajar con un tipo distinto de texto. Algunos trabajan mejor con archivos alfanuméricos, otros con textos de tipo alfabético y otros con archivos con caracteres especiales. La estrategia adoptada al generar estas diferentes versiones de texto permite ejecutar un conjunto de experimentos más amplio.

El proceso anterior se describe de manera gráfica en la siguiente imagen:

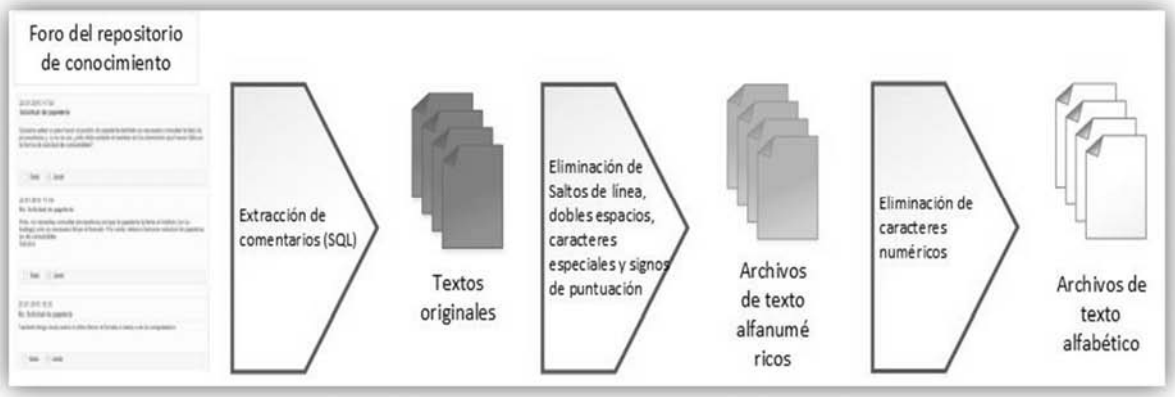


Figura 5-2 Preprocesamiento de los textos.

Fuente: Elaboración propia.

5.2.4. Búsqueda de componentes ontológicos de las secciones del repositorio en los textos

Una vez que se han extraído los comentarios del foro, es necesario procesar automáticamente los textos para reconocer las entidades contenidas en ellos. Las entidades a buscar

corresponden a los elementos de la estructura del repositorio de conocimiento. En otras palabras, buscamos componentes ontológicos de las secciones del repositorio en los textos. Estas entidades se encuentran visibles en la ontología del repositorio creada con anterioridad. Conviene recordar que, para fines de la presente investigación, la experimentación se concentra sobre la sección *administración del laboratorio* del repositorio de conocimiento del GIL.

La siguiente tabla muestra los elementos de la ontología de dicha sección. La primera columna contiene el nombre del elemento que se encuentra en la sección, la segunda muestra las entidades que conforman al elemento, y la tercera muestra la entidad a buscar en el texto. Por ejemplo, si se desea saber si un texto contiene el elemento ontológico “integrante”, lo que realmente se busca en el texto es un “nombre propio” ya que todo integrante tiene un nombre propio. La desventaja de hacer esto es que al encontrar un nombre propio en el texto es difícil identificar si corresponde a un “integrante” (elemento ontológico de “asignación de recursos”) o a un “actor” (elemento ontológico de “procedimiento”), u alguna entidad de otro elemento. En tal caso, al encontrar un nombre propio en el texto, tendremos que considerar que dicho texto contiene todas las entidades que corresponden a un nombre propio.

Elemento	Entidad del elemento	Entidad a buscar
Directorio	Nombre propio	Nombre propio
	Teléfono	Teléfono
	Extensión	Extensión
	Puesto	Puesto
Asignación de recursos	Integrantes	Nombre propio
	Espacios de trabajo	Espacio de trabajo
	Plano	Plano
	Llaves	Clave de llave
	Tarjetas	Clave de tarjeta
Eventos internos	Título del evento	Título
	Contenido	Contenido
	Fecha del evento	Fecha
Procedimientos	Título del procedimiento	Título
	Contenido	Contenido
	Fecha de actualización	Fecha
	Periodicidad	Periodicidad
	Actores	Nombre propio
	Entradas	Nombre de documento
	Salidas	Nombre de documento
Galería	Título de la galería	Título
	Nombre de la galería	Título
	Descripción	Descripción
	Imagen	Imagen
Recursos para la administración	Archivos	Nombre de archivo
	Título del archivo	Título
	Tamaño del archivo	Tamaño
Cómputo	Título	Título
	Contenido	Contenido
Foro	Publicador	Nombre propio
	Título	Título
	Mensaje	Mensaje
	Fecha de publicación	Fecha
Tesis	Metodología	Nombre de sección de tesis
	Capítulos	Nombre de sección de tesis
	Fecha	Fecha
	Hipótesis	Nombre de sección de tesis
	Preguntas de investigación	Nombre de sección de tesis
	Objetivos	Nombre de sección de tesis
	Tutor	Nombre propio
	Autor	Nombre propio
	Grado	Grado académico
	Título	Título
	Resultados	Nombre de sección de tesis
	Metodología	Nombre de sección de tesis
	Experimentos	Nombre de sección de tesis
Expedientes	Archivos	Nombre de archivo de expediente
	Título	Título
	Tamaño	Tamaño

Tabla 5-2 Componentes ontológicos (entidades) de Administración del laboratorio

Quitando la sección *foro* (los textos ya están en esa sección, de ahí se extraen) y removiendo elementos duplicados de la tercera columna tenemos la siguiente lista con las entidades que necesitamos reconocer dentro de los textos extraídos:

1. Nombre propio
2. Teléfono
3. Extensión
4. Puesto
5. Espacio de trabajo
6. Plano
7. Clave de llave
8. Clave de tarjeta
9. Título
10. Contenido
11. Fecha
12. Periodicidad
13. Nombre de documento
14. Descripción
15. Imagen
16. Nombre de archivo
17. Tamaño
18. Grado académico
19. Nombre de sección de tesis
20. Nombre de archivo de expediente

Para poder reconocer estas entidades en los textos extraídos se ha recurrido al uso de etiquetado Part Of Speech (POS), expresiones regulares y gazeteers (listas). A continuación se presenta una breve descripción de estos y se muestra la forma en que fueron aprovechados en la presente investigación.

Etiquetado POS

Para encontrar nombres propios y entidades nombradas en los textos se utilizó una herramienta llamada FreeLing. Esta herramienta recibe como entrada un archivo de texto y entrega como salida un nuevo archivo con el texto etiquetado palabra por palabra. Las etiquetas que el sistema coloca nos permitieron conocer la siguiente información:

1. Categoría gramatical a la que pertenece cada palabra. Ejemplo: *asista* / VMSP3S0 (Verbo, principal, subjuntivo, tiempo presente, tercera persona, singular)
2. Reconocimiento de multi palabras. Palabras que tienen un significado propio cuando están juntas. Por ejemplo: *Tal vez.*
3. Encontrar entidades nombradas. Elementos que son una sola entidad aunque su nombre esté compuesto por varias palabras. Por ejemplo, *Universidad Nacional Autónoma de México.*

Cabe precisar que la herramienta provee más información e incluye más funcionalidades, pero para nuestro trabajo aprovechamos únicamente algunas de ellas. En especial se aprovechó el reconocimiento de nombres propios, estos son marcados por la herramienta con la etiqueta *NP*.

A continuación se muestra un ejemplo real de un texto extraído del foro del repositorio de conocimiento y su etiquetado correspondiente una vez que fue procesado por la herramienta FreeLing.

Texto original:

cuando se realiza un proyecto duro la secretaría administrativa tiene que generar una factura. este trámite es con Brenda Figueroa.

Texto etiquetado:

cuando cuando CS 0.985595
se se P00CN000 0.465639
realiza realizar VMIP3S0 0.994868
un uno DI0MS0 1
proyecto proyecto NCMS000 0.986111
duro duro AQ0MS0 0.975904
la el DA0FS0 0.972269
secretaria secretaria NCFS000 0.648934

administrativa administrativo AQ0FS0 0.53725
tiene tener VMIP3S0 1
que que CS 0.437483
generar generar VMN0000 1
una uno DI0FS0 0.972376
factura factura NCFS000 0.846947
este este DD0MS0 0.960092
trámite trámite NCMS000 1
es ser VSIP3S0 1
con con SPS00 1
Brenda_Figueroa brenda_figueroa NP00000 1

Cada línea está compuesta por la palabra original (tal cual aparece en el texto procesado), seguida del lema de la palabra, esto es, la palabra en su forma general; luego aparece una etiqueta que indica la categoría gramatical a la que probablemente pertenece, de acuerdo al estándar EAGLES[4]; después se encuentra la probabilidad de que la palabra pertenezca a la categoría gramatical señalada.

Expresiones regulares

El reconocimiento de números telefónicos, extensiones telefónicas, fechas, espacios de trabajo, claves de llave y de tarjeta, y tamaños de archivos se realizó mediante expresiones regulares (REGEX). Estas expresiones son un conjunto de símbolos que representan el patrón que ha de buscarse en una cadena de caracteres, es decir, en un texto. Dicho patrón indica una serie de condiciones que deberá cumplir secuencialmente un texto para ser reconocido como un elemento compatible con la expresión de búsqueda.

Por ejemplo, dado el siguiente texto:

Se acabó la punta de mi lápiz, no puedo escribir ni un punto más. No sirve mi sacapuntas. Necesito un sacapuntas nuevo.

[4]El Expert Advisory Group on Language Engineering Standards (EAGLES) es un grupo que se encarga de proveer estándares para diversas tareas, entre ellas la anotación de lexicones. Las etiquetas EAGLES se utilizan frecuentemente para la anotación morfosintáctica.
Los detalles de estas etiquetas pueden consultarse en la siguiente dirección electrónica: http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Si se usa la expresión regular '[Pp]unt[ao]' se estará indicando que se desea buscar la cadena “P” o “p” seguida de “unt”, seguida de una letra “a” u “o”. De manera que, en el texto se encontrarán cuatro palabras que contienen las cadenas de caracteres que cumplen con el patrón de búsqueda:

- ***punta***
- ***punto***
- *Saca**punta**s*.
- *Saca**punta**s*

En cambio, si se usa la expresión '[Pp]unt[ao]\s' se buscará la misma cadena que el ejemplo anterior pero seguida de un espacio en blanco. Entonces, en el texto solamente existirán dos palabras que contienen la cadena buscada:

- ***punta***
- ***punto***

Para facilitar la comprensión de las expresiones regulares que se mostrarán adelante, se entrega a continuación una lista de los elementos utilizados y su funcionalidad.

Elemento	**Funcionalidad**
[ ]	Representa un conjunto o rango de elementos
( )	Permite agrupar elementos
\b	Frontera de la cadena
\s	Espacio en blanco
?	Convierte en opcional la condición que se encuentre a la izquierda de su posición
|	Operador lógico OR

Las expresiones que se desarrollaron en este trabajo para encontrar los elementos mencionados son las siguientes:

- Búsqueda de teléfonos:

'\b([1-9][0-9][-\s]?[0-9][0-9][-\s]?[0-9][0-9][-\s]?[0-9][0-9])\b'

Ejemplos de cadenas coincidentes con el patrón de búsqueda:

56 23 36 00

56-08-16-23

5683 2283

- Búsqueda de extensiones:

' ([Ee]xt.?\s|[eE]xtension|[Ee]xtensión)'

Ejemplos de cadenas coincidentes con el patrón de búsqueda:

Ext.
ext
extension
Extensión

- Búsqueda de fechas:

'([0-3]?[0-9]\sde\s([eE]nero|[Ff]ebrero|[Mm]arzo|[Aa]bril|[Mm]ayo|[Jj]unio|[Jj]ulio|[Aa]gosto|[Ss]epriembre|[Oo]ctubre|[Nn]oviembre|[Dd]iciembre))'

'((([0-2]?[0-9])|([3][0-1]))[-/]((([0]?[1-9])|([1][0-2]))|([eE]nero|[Ff]ebrero|[Mm]arzo|[Aa]bril|[Mm]ayo|[Jj]unio|[Jj]ulio|[Aa]gosto|[Ss]epriembre|[Oo]ctubre|[Nn]oviembre|[Dd]iciembre)|([eE]ne|[Ff]eb|[Mm]ar|[Aa]br|[Mm]ay|[Jj]un|[Jj]ul|[Aa]go|[Ss]ep|[Oo]ct|[Nn]ov|[Dd]ic))[-/][1-2][0-9][0-9][0-9])'

Ejemplos de cadenas coincidentes con el patrón de búsqueda:

12-11-2014
06/marzo/1910
30-Noviembre-2002
04/Dic/1983

- Búsqueda de claves de tarjeta:

'(\d{3}[-]\d{5}[-\s]\d{8}[-]\d)'

Ejemplos de cadenas coincidentes con el patrón de búsqueda:

038-12993 11304915-1

101-39875-65894521-2

- Búsqueda de claves de llave y espacios de trabajo (ambos elementos se identifican con el mismo código):

'(([a-d]|[A-D])[-][0-1][0-9])'

Ejemplos de cadenas coincidentes con el patrón de búsqueda:

C-02
d-16

- Búsqueda de tamaños de archivo:

'(\d+[\.][\d+][\s]?[KkMmGgTtPp][Bb]])'

'(\d+[\s]?[KkMmGgTtPp][Bb])'
Ejemplos de cadenas coincidentes con el patrón de búsqueda:

236.58 Mb
6152.2KB
23785 Gb
5TB

Gazeteers

Las expresiones regulares permiten encontrar cadenas de texto que concuerdan con un patrón determinado. Sin embargo, algunos de los elementos que necesitamos reconocer no tienen algún patrón predecible en el texto, de manera que no es posible utilizar este mecanismo de reconocimiento para dichos elementos.

Para superar esto se recurrió a la búsqueda directa del conjunto de palabras que dan nombre a las instancias del objeto a buscar. Por ejemplo, para el objeto *puesto de trabajo* las instancias podrían ser *director de recursos humanos*, *encargado de bodega*, etc. Estos

últimos elementos formaron la lista de cadenas de texto a buscar dentro de los comentarios extraídos. El conjunto de listas creadas conforman los Gazeteers.

Cabe mencionar que la lista de palabras puede ser considerablemente extensa, pero está limitada por los elementos existentes en el dominio en estudio. Por ejemplo, para el caso de *puesto de trabajo*, la lista se limita a los nombres de los puestos de trabajo existentes en la institución a la que pertenece el grupo de investigación que sirve como caso de estudio.

A manera de ejemplo se muestra el gazeteer generado para reconocer la *periodicidad*:

- diaria
- semanal
- mensual
- bimestral
- trimestral
- cuatrimestral
- semestral
- anual
- bianual
- a demanda

La lista de palabras se realizó de manera manual para algunos elementos y para otros se obtuvo de manera automática mediante consultas SQL. Esto último responde a la intuición de que en los textos se podrían encontrar palabras que nombran directamente a elementos del repositorio de conocimiento; a los cuales se está haciendo una aportación, crítica o consulta. Por ejemplo: “ya no trabaja aquí la contadora general, pero algunos procedimientos como el de solicitud de papelería siguen mostrando su nombre”. Por lo tanto, la lista de palabras generada automáticamente debe contener todos los títulos de los procedimientos existentes en el repositorio, incluyendo *solicitud de papelería.*

Las listas hechas manualmente permitieron reconocer los siguientes elementos:

1. Puestos de trabajo
2. Periodicidad
3. Nombre de archivo
4. Grado académico
5. Nombres de sección de tesis

6. Nombres de archivos de expediente
7. Nombres de documentos de entradas y salidas de procedimientos

Las listas obtenidas automáticamente mediante consultas SQL a la base de datos del manejador de contenido donde se aloja el repositorio sirvieron para reconocer títulos.

La lista completa de gazeteers utilizados en este trabajo, realizados manualmente, puede consultase en el anexo 8. Las listas creadas automáticamente no se muestran porque son creadas dinámicamente y su contenido cambia constantemente.

El procesamiento de los textos, con base en lo descrito anteriormente, se expone de manera gráfica en la siguiente imagen:

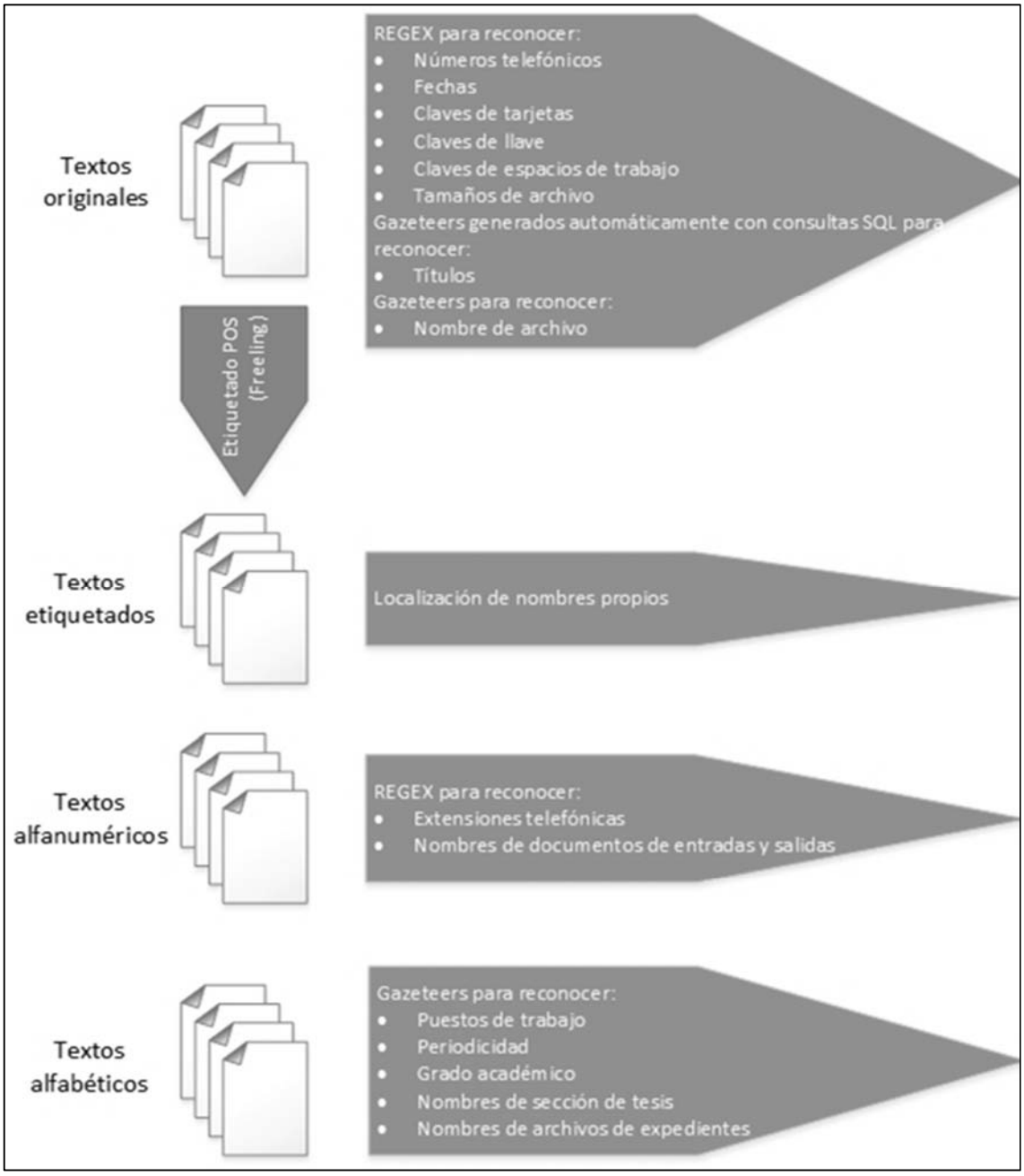


Figura 5-3 Reconocimiento de componentes ontológicos

Fuente: Elaboración propia

A continuación se muestra el ejemplo de un texto real extraído del repositorio de conocimiento y la captura de pantalla de la salida que entrega el sistema que se elaboró para implementar los mecanismos de reconocimiento descritos anteriormente.

Texto original (36.txt):

El contacto responsable en conacyt para becas de proyectos de Básicas es la Bióloga Patricia Ojeda Subdirectora de control de proyectos de investigación Tel. 53227700 ext. 6610

Salida del sistema:

```
**********************************************************
ARCHIVO:  36.txt
Teléfonos encontrados:  si
Nombres propios encontrados:  si
Extensiones encontradas:  si
Fechas encontradas:  no
Tarjetas encontradas:  no
Llaves encontradas:  no
Espacios de trabajo encontrados:  no
Tamaños de archivo encontrados:  no
Puestos encontrados:  no
Periodicidades encontradas:  no
Nombres de archivo encontrados:  no
Grados académicos encontrados:  no
Secciones de tesis encontrados:  no
Nombres de archivos de expediente encontrados:  no
Títulos encontrados:  no
Nombres de documentos encontrados:  no
**********************************************************
```

En este ejemplo, el teléfono encontrado es el *53227700*, los nombres propios reconocidos son *Básicas* (erróneamente) *y Patricia Ojeda*, y la extensión localizada es *6610.*

Como puede observarse en el ejemplo, las estrategias de reconocimiento que se adoptaron en esta investigación y que se han descrito anteriormente, tienen errores de tipo falso positivo y falso negativo. Los primeros se refieren a elementos que el sistema reconoce como un elemento buscado, sin serlo realmente. Por ejemplo, la palabra “Básicas” fue reconocida por el sistema como un nombre propio. Los segundos, al contrario, se refieren a una omisión del sistema cuando este no es capaz de reconocer un elemento que sí está presente en el texto. Por ejemplo, el texto contiene la cadena “Subdirectora de control de proyectos de investigación” pero el sistema no identifica que se trata de un puesto de trabajo.

Sin embargo, la finalidad del presente trabajo no es optimizar el reconocimiento de dichos elementos, sino observar la posibilidad, viabilidad y oportunidad de aplicar técnicas y métodos de las tecnologías del lenguaje para procesar automáticamente el conocimiento explicitado en las herramientas de los repositorios de conocimiento.

Además, como ya se ha mencionado anteriormente, el alcance del presente trabajo se limita a la experimentación con el repositorio de conocimiento del Grupo de Ingeniería Lingüística (Caso de estudio). Por esta razón, las estrategias de extracción y procesamiento del conocimiento fueron diseñadas específicamente para el dominio en cuestión, sin pretender que estas sean usables directamente en otros repositorios de conocimiento. Sin embargo, es posible hacer modificaciones relativamente sencillas para que estas puedan utilizarse en otros dominios.

Elementos no reconocidos

En una sección anterior se mostró una lista de elementos que deben buscarse en los textos extraídos. No obstante, algunos de estos elementos son muy difíciles de reconocer con las estrategias de reconocimiento adoptadas en este trabajo, entre los que se encuentran “plano”, “imagen” y “descripción”. Esto se debe a que sus características confieren una complejidad notablemente mayor a los mecanismos que pretenden distinguirlos. Por ejemplo, el elemento “descripción” está compuesto por texto que es indistinguible de otros textos, de manera que no es posible saber si una cadena corresponde a un texto cualquiera o a una descripción de algo.

Por lo tanto, para efectos de la presente investigación se omiten dichos elementos y se trabajará únicamente con aquellos que son identificables con las estrategias de reconocimiento diseñadas. Esto se hace teniendo en cuenta que los elementos reconocibles por las estrategias ideadas conforman el 80% del total de tipos de elementos existentes en la sección del repositorio de conocimiento elegida para hacer la experimentación.

5.2.5. Categorización automática del conocimiento

Una vez que se conocen los elementos que están contenidos en cada texto, es necesario compararlos con aquellos elementos que integran cada sección de la estructura del repositorio

de conocimiento. La hipótesis que se intuye es que a mayor número de elementos coincidentes entre un texto y un elemento de la estructura del repositorio, mayor probabilidad existe de que dicho texto deba pertenecer al elemento mencionado.

Para tal efecto, se programó un sistema que genere automáticamente unas estructuras que nos permitan hacer dicha comparación. Con este sistema, inicialmente generamos una estructura tabular (matriz) donde las columnas representan las entidades existentes que configuran los elementos de las secciones del repositorio de conocimiento (E_n), los renglones representan los textos extraídos, y las intersecciones entre columnas y renglones (v) indican, a través de una etiqueta, la aparición (1) o no aparición (0) del elemento en el texto. A esta matriz se le llamó *matriz de apariciones*. Suele utilizarse el término vector para referirse a un renglón de la matriz.

La estructura de la matriz de apariciones se muestra a continuación.

Número_de_texto_extraído	E1	E2	E3	E4	…	E14	E15	E16
1.txt	*v*	*v*	*v*	*v*	*v*	*v*	*v*	*v*
2.txt	*v*	*v*	*v*	*v*	*v*	*v*	*v*	*v*
3.txt	*v*	*v*	*v*	*v*	*v*	*v*	*v*	*v*
…	*v*	*v*	*v*	*v*	*v*	*v*	*v*	*v*
n.txt	*v*	*v*	*v*	*v*	*v*	*v*	*v*	*v*

Tabla 5-3 Estructura de la matriz de apariciones

Fuente: Elaboración propia

Donde[5]:

E1 = Teléfono
E2 = Nombre propio
E3 = Extensión
E4 = Fecha
E5 = Clave de tarjeta
E6 = Clave de llave
E7 = Espacio de trabajo
E8 = Tamaño de archivo
E9 = Puesto

[5] En la matriz generada por el programa codificado para el procesamiento de los textos, los encabezados de las columnas no tienen acentos ni espacios, y las mayúsculas se usan para indicar el inicio de la segunda o tercera palabra que conforma el nombre del elemento. Por ejemplo: nombreDeArchivo

E10 = Periodicidad
E11= Nombre de archivo
E12 = Grado académico
E13 = Sección de tesis
E14 = Archivo de expediente
E15 = Título
E16 = Nombre de documento
$v \in \{0,1\}$

Posteriormente, el sistema programado lee un renglón de la matriz y compara los elementos que aparecen en este contra los que existen en cada elemento de las secciones del repositorio de conocimiento. Esta última información se obtiene de la ontología que se creó anteriormente para modelar el repositorio de conocimiento. Al hacer esto, el sistema cuenta el número de entidades coincidentes y calcula la probabilidad de que este texto (representado por el renglón que se está procesando) tenga conocimiento que pertenece a cada sección del repositorio. Esta probabilidad se calcula a través de una división, donde el dividendo es el número de entidades coincidentes y el divisor es el número total de entidades del elemento de la sección del repositorio con el que se está comparando en ese momento.

$$probabilidad\_de\_pertenencia = \frac{Número\_de\_entidades\_coincidentes}{Número\_total\_de\_entidades\_del\_elemento}$$

Al tomar un renglón (que representa a un texto) y comparado con cada elemento de la estructura del repositorio, se generarán k valores que indican las probabilidades de que ese renglón esté relacionado con cada elemento, donde *k* es el número de elementos de la sección del repositorio donde se hace la experimentación.

Para ilustrar lo descrito anteriormente se muestra la comparación entre un renglón y algunos elementos de la sección del repositorio.

Entidades encontradas en el comentario "36.txt"	Entidades del elemento "Directorio"
• Teléfono • Nombre propio • Extensión telefónica	• Teléfono • Nombre propio • Extensión telefónica • Puesto Número total de elementos = 4

Número de coincidencias = 3
Probabilidad_ de_pertenencia= (3 / 4) = 0.75

Entidades encontradas en el comentario “36.txt”	Entidades del elemento “Asignación de recursos”
• Teléfono • Nombre propio • Extensión telefónica	• Nombre propio • Espacio de trabajo • Clave de llave • Clave de tarjeta Número total de elementos = 4

Número de coincidencias = 1

Probabilidad_ de_pertenencia= (1 / 4) = 0.25

Entidades encontradas en el comentario “36.txt”	Entidades del elemento “Eventos internos”
• Teléfono • Nombre propio • Extensión telefónica	• Título • Fecha Número total de elementos = 2

Número de coincidencias = 0

Probabilidad_ de_pertenencia= (0 / 2) = 0

Este proceso se repite con todos los renglones de la matriz hasta procesarlos todos. Los valores que se van encontrando se guardan en una nueva matriz donde las columnas representan a los elementos de la sección del repositorio donde se hace la experimentación (A_n), los renglones representan cada comentario extraído, y las intersecciones (*p*) guardan la probabilidad de que el conocimiento contenido en el comentario extraído pertenezca o se relacione con el elemento representado en la columna donde se encuentre. A esta matriz se le llamó *matriz de probabilidad de pertenencia.*

La estructura de la matriz de probabilidad de pertenencia se muestra a continuación.

Número_de_texto_extraído	A1	A2	A3	A4	A5	A6	A7	A8	A9
1.txt	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
2.txt	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
3.txt	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
…	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*
n.txt	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*	*p*

Tabla 5-4 Estructura de la matriz de probabilidad de pertenencia

Fuente: Elaboración propia

Donde[6]:

A1 = Directorio
A2 = Asignación de recursos
A3 = Eventos internos
A4 = Expedientes
A5 = Procedimientos
A6 = Galería
A7 = Recursos para la administración
A8 = Cómputo
A9 = Tesis
$p \in \mathbb{R} : 0 < p < 1$

A manera de ejemplo se muestra el renglón que se genera en la matriz de probabilidades de pertenencia tras el procesamiento del texto "36.txt" (mostrado arriba).

Número_de_texto_extraído	A1	A2	A3	A4	A5	A6	A7	A8	A9
36.txt	**0.75**	0.25	0	0	0.2	0	0	0	0.2

En el ejemplo puede observarse que el texto "36.txt" tiene una mayor probabilidad de tener contenido relacionado a la sección "Directorio" (A1).

Finalmente, el sistema crea una matriz de dos columnas que indica la sección con la que el contenido de cada texto tiene una mayor probabilidad de relación. Esta matriz fue nombrada *matriz de predicciones*. La primera columna contiene el nombre con el que se identifica a cada texto extraído y la segunda indica el elemento de la sección del repositorio con el que tiene relación.

La estructura de la matriz de predicciones se muestra a continuación.

Número_de_texto_extraído	C

Tabla 5-5 Estructura de la matriz de predicciones

Fuente: Elaboración propia

[6] Al igual que la matriz de apariciones, los encabezados de las columnas no tienen acentos ni espacios, y las mayúsculas se usan para indicar el inicio de la segunda o tercera palabra que conforma el nombre del elemento. Por ejemplo: asignacionDeRecursos

Donde:

C ∈ {A1, A2, A3, A4, A5, A6, A7, A8, A9, *Null*}

En cada texto (renglón) se asigna la clase donde el valor de la probabilidad es más alta (columna), siempre y cuando este valor sea mayor que un umbral previamente definido.

Para ejemplificar el contenido de la matriz de predicciones, se muestra a continuación el renglón generado para el texto 36.txt.

36.txt	A1 (Directorio)

En caso de no superar el umbral, la clase que se asigna es *Null*. La finalidad de esto último es permitir que el sistema asigne la clase *Null* cuando encuentra muy pocos elementos en los textos; esto sucede cuando el texto es muy corto o contiene un discurso que no tiene relación con ninguna sección del repositorio. Un ejemplo de este último caso es el siguiente.

Texto 5.txt

A mi me gusta la idea de Wall-e o el gigante de acero, pues ambas tratan el tema de robtos que hablan, lo que tiene que ver con PLN

5.txt	*Null*

5.2.6. Evaluación e inserción del conocimiento en las secciones del repositorio

Como pudo observarse, el sistema categoriza el conocimiento de cada texto del foro en los elementos de las secciones del repositorio y genera una matriz de predicciones que indica el lugar más probable al que pertenece cada texto. Sin embargo, existe en cada predicción un grado de incertidumbre que obliga a que un humano evalúe la pertinencia de enviar cierto conocimiento a un espacio del repositorio o a otro.

El presente trabajo tiene la intención de que el sistema auxilie el proceso de organización del conocimiento de los repositorios, pero no pretende eliminar el trabajo

humano para decidir enviar o no el conocimiento a alguna sección del repositorio de conocimiento. Tampoco pretende insertar automáticamente el conocimiento en la sección definida por la categorización. Por esta razón, en la metodología que se diseña, se utiliza el término organización semiautomática del conocimiento.

El trabajo futuro podría dirigirse a desarrollar un sistema que utilizando la matriz de predicción inserte el conocimiento en el lugar que corresponda a la categorización realizada. Sin embargo, como ya se ha dicho, ese trabajo queda fuera de los alcances de la presente tesis.

5.2.7. Resultados y evaluación

Para medir la eficiencia que tiene el método propuesto para relacionar (categorizar) un comentario con un elemento de una sección del repositorio, se hizo una categorización manual de los comentarios y se comparó contra la categorización automática que genera el método. Esto permitió observar los errores y aciertos que obtiene el sistema.

Para tener una categorización manual objetiva se mostraron los textos a tres personas, de manera individual, y se les solicitó que categorizaran cada texto en las secciones del repositorio o en la clase *Null*, considerando únicamente el contenido del texto. Posteriormente, se conciliaron las tres categorizaciones y se creó una matriz para guardar la categorización de cada texto. Esta matriz tiene la misma estructura que la matriz de predicciones y se le llamó matriz de precategorizados.

El consenso de las tres categorizaciones siguió las siguientes reglas:

a. Si las tres categorizaciones coinciden en la misma clase, entonces esa será la clase consensada.
b. Si dos de las tres categorizaciones coinciden en la misma clase, entonces esa clase se tomará como clase consensada.
c. Si las clases son diferentes en las tres categorizaciones, entonces se solicita a un cuarto individuo que clasifique ese texto en alguna de las clases propuestas por las otras tres categorizaciones. Esta cuarta categorización define la clase que se tomará como consensada.

Una vez que se tienen las matrices de predicciones y precategorizados, un programa que se codificó para tal efecto las compara para encontrar coincidencias y diferencias. Con estos datos el mismo programa calcula las siguientes tres medidas de eficiencia por cada clase:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F1\ score = \frac{2\ (precision)(recall)}{precision + recall}$$

Donde:

- *tp* = verdaderos positivos = elementos clasificados correctamente en determinada clase
- *fp* = falsos positivos = elementos clasificados incorrectamente en determinada clase
- *fn* = falsos negativos = elementos que debieron clasificarse en determinada clase pero que fueron clasificados en otra

A continuación se describe la interpretación de las medidas de eficiencia utilizadas.

- *Precision*: Del número total de elementos clasificados en esa clase, ¿cuántos realmente debían ser clasificados en ella?
- *Recall*: De todos los textos existentes que pertenecen a una clase, ¿cuántos encontró el sistema? Y ¿cuántos faltaron por encontrar?
- *F1-score* = Medida armónica que balancea la *precision* y el *recall*.

Al terminar de obtener las medidas anteriores se calculan las siguientes medidas para observar la eficiencia general del sistema.

$$Precision\_promedio = \frac{\sum_{i=0}^{k-1} presicion\ (i)}{k}$$

$$Recall\_promedio = \frac{\sum_{i=0}^{k-1} recall\ (i)}{k}$$

$$F1score\_promedio = \frac{\sum_{i=0}^{k-1} F1score\ (i)}{k}$$

$$Accuracy = \frac{\sum_{i=0}^{k-1} tp\ (i)}{nt}$$

Donde:

- *k* = número de clases
- *nt* = número total de textos

Además, el programa genera un arreglo bidimensional cuadrado, llamado matriz de confusión. Esta matriz sirve para facilitar la observación de los errores y aciertos en las predicciones realizadas por el sistema. Esta matriz cuadrada tiene un tamaño de k*k dimensiones donde k es el número de clases donde pueden clasificarse los textos. El eje de las *y* (vertical) indica la clase que se esperaba tener para un determinado elemento y el eje de las *x* (horizontal) muestra la clase a la que fue asignado dicho elemento por el sistema. Por lo tanto, en la diagonal de la matriz se encontrarán los elementos que el sistema clasificó correctamente, ya que se encontrarán en la posición P[*i*][*j*] donde *i=j*, esto quiere decir que la clase esperada para cierto elemento es la misma clase encontrada por el sistema.

Los elementos que quedan fuera de la diagonal serán aquellos elementos para los que se esperaba una clase *i* y el sistema le asignó la clase *j*. Por eso la posición donde se encuentra es P[*i*][*j*] donde *i≠j*.

A continuación se muestra la matriz de confusión generada tras la comparación de los resultados del primer experimento así como sus métricas de eficiencia:

```
[ 4.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  1.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  5.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  3.  0.  0.  0.  0.  6.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  1.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.]
[  0.   0.   0.   0.   0.   0.   0.   0.   0.  22.]
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4
1	0.00	0.00	0.00	1
2	0.00	0.00	0.00	5
4	1.00	0.33	0.50	9
6	0.00	0.00	0.00	1
9	0.63	1.00	0.77	22
avg / total	0.64	0.69	0.61	42

Accuracy= 29 /42 = 0.69

En la matriz de confusión podemos observar que la mayoría de los textos correctamente categorizados pertenecen a la categoría *Null* (22). Esto tiene sentido, ya que la mayoría de los textos publicados en el foro administrativo del repositorio del GIL no aportan directamente nuevo conocimiento, sino que tienen la intención de solicitar alguna información a los otros usuarios, provocar el debate, replicar algún comentario o exponer nuevas ideas. Solamente un número reducido de textos agrega conocimiento novedoso o actualiza al ya existente. En consecuencia, el alto número de comentarios categorizados en dicha categoría indica que el sistema trabaja de manera adecuada.

Para incrementar la publicación de textos que aportaran nuevo conocimiento y, en consecuencia, tener más elementos pertenecientes a clases distintas a *Null* y experimentar con ellas, se creó un nuevo hilo en el foro administrativo llamado “nuevo conocimiento”. La intención de este hilo es que los usuarios del repositorio pudieran crear en él nuevos comentarios (post) que tengan la intención exclusiva de aportar nuevo conocimiento, o actualizarlo. Se aprovechó el seminario semanal del GIL para promover su uso.

La estrategia anterior funcionó relativamente bien, pues fomentó la creación de textos con esas características. Sin embargo, el número de textos creados a partir de esta estrategia es limitado. Esto se debe a la dinámica propia de la creación de conocimiento de la organización, pues el conocimiento administrativo del GIL, que apoya a sus procesos productivos, es notoriamente inferior al conocimiento que este genera como producto de dichos procesos. Además, el número de integrantes que tiene la organización es pequeño, y la utilización que estos hacen del foro apenas comienza a incrementarse. En consecuencia, la limitación en el número de comentarios realizados en el foro con estas características es inevitable.

A pesar de esto, en la matriz de confusión puede observarse que existen varios elementos que son correctamente categorizados en las distintas clases existentes. La línea diagonal de la matriz contiene 7 elementos que corresponden a las categorizaciones distintas a la clase *Null*, realizadas correctamente por el sistema. Dichas categorizaciones se realizaron para las clases *Directorio* y *Procedimientos*.

Por otro lado, se puede ver que existen 13 elementos que fueron incorrectamente categorizados en la clase *Null*. Esto indica que el sistema no logra relacionar dichos textos con las secciones del repositorio y los categoriza a la clase *Null*. Por ejemplo, el texto "22.txt" fue categorizado a la clase *Null* pero al leerlo observamos que debió ser categorizado a la clase *Procedimientos.*

Texto (22.txt):

La apertura de proyectos ya no es a través del sistema. Hay que llenar un formato con Macario.

Otro ejemplo es el texto "27.txt" que fue categorizado a la clase *Null* a pesar de contener información posiblemente relacionada con las clases *Eventos Internos*, *Directorio* y *Procedimientos*.

Texto (27.txt):

Para realizar cualquier evento académico existe en el iingen un departamento encargado de la logística "coordinación de eventos"

Para solucionar este problema es posible mejorar las herramientas desarrolladas para hacer el reconocimiento de los componentes ontológicos en los textos, o bien, modelar una ontología que modele con más detalle los elementos del repositorio de conocimiento del GIL. Sin embargo, estas mejoras quedan fuera del alcance de este trabajo, pues el objetivo que este persigue es únicamente mostrar la posibilidad de integrar las tecnologías del lenguaje en una metodología que permita auxiliar la organización de los repositorios de conocimiento mediante un proceso semiautomático. Por lo tanto, estas mejoras se mencionan a manera de trabajo futuro.

5.3. Metodología propuesta

Después de realizar la experimentación descrita en las secciones previas y observar los resultados obtenidos, se observa que los pasos realizados operan correctamente para procesar automáticamente el conocimiento explicitado en el foro del repositorio de conocimiento y permiten categorizarlo (también de manera automática) con un nivel aceptable de eficiencia. Dicha categorización proporciona información valiosa para apoyar la organización de los repositorios de conocimiento.

Por esta razón, la serie de pasos descritos con anterioridad se propone como una metodología para organizar semiautomáticamente el conocimiento de los repositorios. Dicha metodología consiste en aprovechar la taxonomía del conocimiento que se utilizó en la creación de la estructura del repositorio de conocimiento, para modelar las secciones de dicho repositorio a través de ontologías. Esta ontología describirá los elementos que deben buscarse en los textos en una etapa posterior.

De manera paralela, se extraen los textos de las herramientas de la administración del conocimiento (integradas en el repositorio) y se aplica un preprocesamiento que consiste en crear tres archivos de texto por cada uno de los comentarios. El primero contiene el comentario original. El segundo guarda el texto alfanumérico sin saltos de línea, dobles espacios y caracteres especiales como signos de puntuación, guiones, etc. El tercero es similar al anterior, pero excluye caracteres numéricos.

Posteriormente, se buscan los componentes ontológicos descritos en la ontología creada en un paso previo. Para tal efecto es posible aprovechar varias técnicas, modelos y

herramientas pertenecientes a las tecnologías del lenguaje, como el reconocimiento de entidades que forma parte de la extracción de información.

Luego, se calcula el grado de relación que tiene un texto con cada sección del repositorio de conocimiento. Para tal efecto se realiza el siguiente cálculo:

$$probabilidad\_de\_pertenencia = \frac{Número\_de\_entidades\_coincidentes}{Número\_total\_de\_entidades\_del\_elemento}$$

Finalmente, un humano evalúa la pertinencia de colocar el conocimiento en la sección categorizada por el sistema y, en caso de serlo, inserta manualmente dicho conocimiento en la sección determinada.

A continuación se resume de manera gráfica este conjunto de pasos.

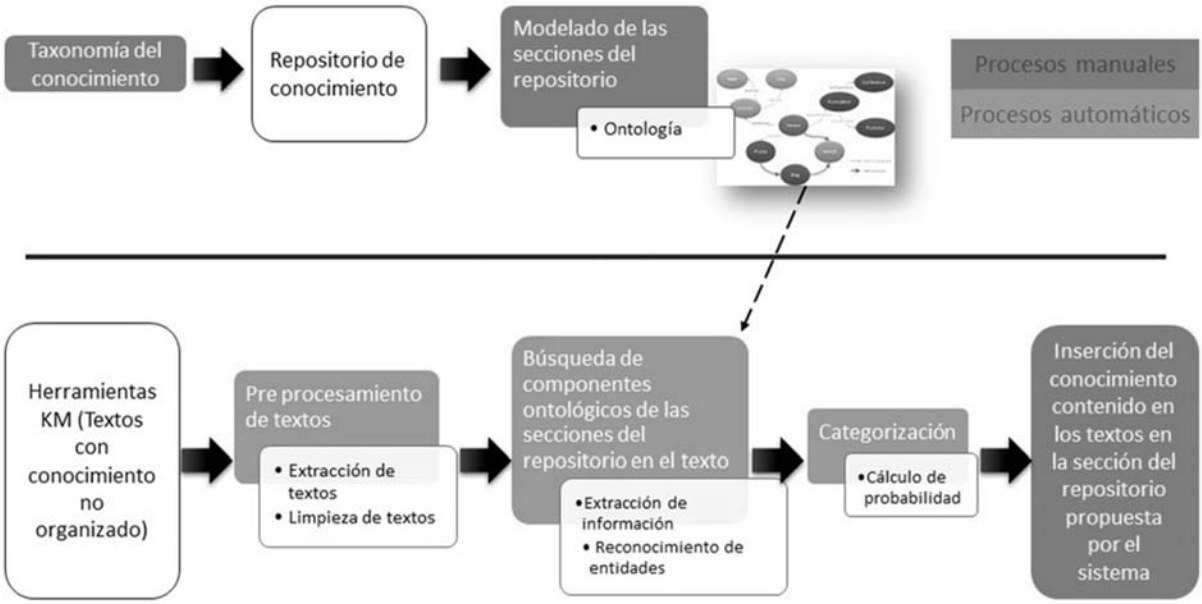


Figura 5-4 Metodología propuesta para organizar semiautomáticamente nuevo conocimiento

CAPÍTULO 6. Conclusiones

El presente trabajo de investigación permite concluir:

Las tecnologías del lenguaje pueden ser aprovechadas para explotar los repositorios de conocimiento de una manera más eficiente. Esto debido a que se pudo confirmar que el conocimiento explicitado en las herramientas de la administración del conocimiento basadas en TI puede ser procesado de una manera más eficiente por las tecnologías del lenguaje.

Es posible organizar semiautomáticamente el conocimiento explicitado en las herramientas de la administración del conocimiento integradas en los repositorios de conocimiento, aprovechando las tecnologías del lenguaje, mediante la metodología diseñada en el presente trabajo de investigación.

La administración del conocimiento puede ser implementada en organizaciones dedicadas a la investigación haciendo ligeras adaptaciones a las metodologías de implementación y las herramientas que estas integran, propuestas por diversos autores. Las mayores adaptaciones corresponden a la modificación del léxico (palabras) utilizado en las herramientas, para contextualizarlo al ámbito académico y de investigación.

Existen muchas herramientas de la administración del conocimiento que permiten identificar, crear, almacenar, compartir y aplicar el conocimiento de las organizaciones. La elección de las que se utilizan en una implementación depende de las características propias de la organización donde esta se realice.

El hecho de que la dinámica de trabajo, propia del grupo de investigación, haya generado un número limitado de textos que aportaran nuevo conocimiento en el foro del repositorio, confirma las afirmaciones de diversos autores en relación a que la tecnología habilita y soporta la administración del conocimiento, pero el factor principal son las personas.

6.1. Revisión de objetivos

Se afirma que se cumplieron los objetivos establecidos para el presente trabajo de investigación, dado lo siguiente:

1. Se diseñó una metodología que aprovecha las tecnologías del lenguaje para organizar semiautomáticamente el contenido de repositorios de conocimiento.
2. Se creó un repositorio de conocimiento tras la implementación de la administración del conocimiento en un grupo de investigación, el Grupo de Ingeniería Lingüística.
3. Se aplicaron métodos y técnicas de las tecnologías el lenguaje para explotar de manera más eficiente el repositorio de conocimiento creado.
4. Se observaron y advirtieron las particularidades de la implementación de la administración del conocimiento en un grupo de investigación.

6.2. Revisión de hipótesis

La experimentación realizada permite aceptar la hipótesis planteada al inicio de la investigación, pues demuestra que es posible aplicar tecnologías del lenguaje para procesar el conocimiento explicitado por las herramientas de los repositorios de conocimiento con la finalidad de organizarlo semiautomáticamente.

Además, se confirma que las tecnologías del lenguaje permiten explotar los repositorios de conocimiento de una manera más eficiente.

6.3. Desventajas del método

Se observan las siguientes desventajas en la explotación y organización de los repositorios mediante tecnologías del lenguaje:

- El tiempo requerido para procesar las tesis almacenadas en el repositorio de conocimiento y generar automáticamente los clusters fue alto. Además, el número de documentos que se almacena crece a lo largo del tiempo. De manera que la explotación de los repositorios de conocimiento mediante tecnologías del lenguaje puede requerir de equipo con capacidades considerables de cómputo que soporten el procesamiento de los documentos. Sin embargo, también es posible que existan algoritmos o librerías optimizadas para procesar ágilmente una colección de documentos que no fueron contemplados en esta tesis.
- La evaluación de los clusters generados automáticamente es complicado, pues requiere la interpretación por parte de personas familiarizadas con los temas tratados

en las tesis y cercanas a la producción de las mismas dentro de la organización. Además la agrupación automática de estos elementos podría estar revelando nuevo conocimiento que el evaluador podría ignorar de manera involuntaria.

- La metodología propuesta para la organización semiautomática de los repositorios de conocimiento requiere del modelado de las secciones del repositorio mediante ontologías. Para que la metodología tenga un buen desempeño se debe crear una ontología que tenga un nivel de detalle suficiente para que esta describa adecuadamente los componentes que caracterizan a cada sección del repositorio. Lo anterior requiere de un nivel adecuado de conocimiento sobre creación de ontologías, pero, sobre todo, conocer a detalle la estructura del repositorio de conocimiento.
- Cuando se agrega o modifica la estructura del repositorio de conocimiento, es necesario agregar nuevos elementos a la ontología creada o modificar sus elementos. Este proceso requiere de esfuerzo pues generalmente es un proceso manual (aunque existen trabajos de investigación que buscan desarrollar sistemas de creación automática de ontologías).
- La metodología propuesta aprovecha la taxonomía del conocimiento de la organización. Esta taxonomía es creada generalmente en una etapa previa a la creación del repositorio de conocimiento. Sin embargo, la creación de dicha taxonomía es sugerida, más no obligada, en la implementación de la administración del conocimiento.

6.4. Ventajas del método

Se observan las siguientes ventajas que tienen la explotación y organización de repositorios de conocimiento mediante tecnologías del lenguaje sobre los métodos tradicionales:

- La administración del conocimiento es un proceso continuo de creación, almacenamiento, transmisión y aplicación del conocimiento de la organización, de manera que la asistencia que el método propuesto brinda para organizar los repositorios es valiosa, pues disminuye el esfuerzo y los recursos necesarios para realizar dicha tarea.
- Los clusters generados automáticamente pueden revelar una estructura más eficiente para organizar los documentos procesados, lo que facilitaría el acceso al conocimiento requerido.

- La metodología propuesta para la organización semiautomática de los repositorios de conocimiento es independiente del dominio de aplicación. Esto quiere decir que puede trabajar en repositorios de conocimiento de cualquier institución, sin importar el tamaño, giro comercial o sector al que esta pertenece.
- En la actualidad, la administración del conocimiento es entendida como un trabajo interorganizacional. Estas organizaciones pueden existir en diferentes regiones geográficas comunicándose en diferentes idiomas. Una ventaja de la metodología propuesta es que para procesar textos en diferentes idiomas no es necesario modificar la ontología creada, únicamente se requieren modificaciones a los sistemas de extracción de información desarrollados (autómatas, listas y etiquetado POS).
- A diferencia de otros métodos de clasificación del área de aprendizaje automático, el utilizado en la metodología propuesta para la organización semiautomática de repositorios de conocimiento no requiere de un número grande de textos. Además puede trabajar con textos cortos, como el caso de los textos del foro del repositorio.
- La aplicación de la metodología de organización semiautomática propuesta no requiere de mucha capacidad del equipo de cómputo, y el tiempo que este tarda en procesar los textos es mínimo.

6.5. Trabajo futuro

Se propone el siguiente trabajo futuro para superar los alcances establecidos para el presente trabajo de investigación.

1. Desarrollar un sistema que, utilizando la matriz de predicción creada en este trabajo de investigación, inserte automáticamente el conocimiento en el lugar que corresponda a la categorización realizada.

2. Mejorar el reconocimiento de los componentes ontológicos en los textos, modificando las herramientas creadas en este trabajo o creando nuevas.

3. Generar una ontología que modele con más detalle los elementos del repositorio de conocimiento del GIL para mejorar la categorización del conocimiento.

4. En cuanto se tengan más textos en el foro del repositorio, realizar los mismos experimentos de categorización con un conjunto más grande de textos.

5. Utilizar otros algoritmos de clustering para encontrar el más eficiente para procesar los documentos de un repositorio de conocimiento.

6. Replicar la experimentación en repositorios de conocimiento de entidades de naturaleza diferente a un grupo de investigación para observar el comportamiento del método propuesto.

6.6. Observaciones finales

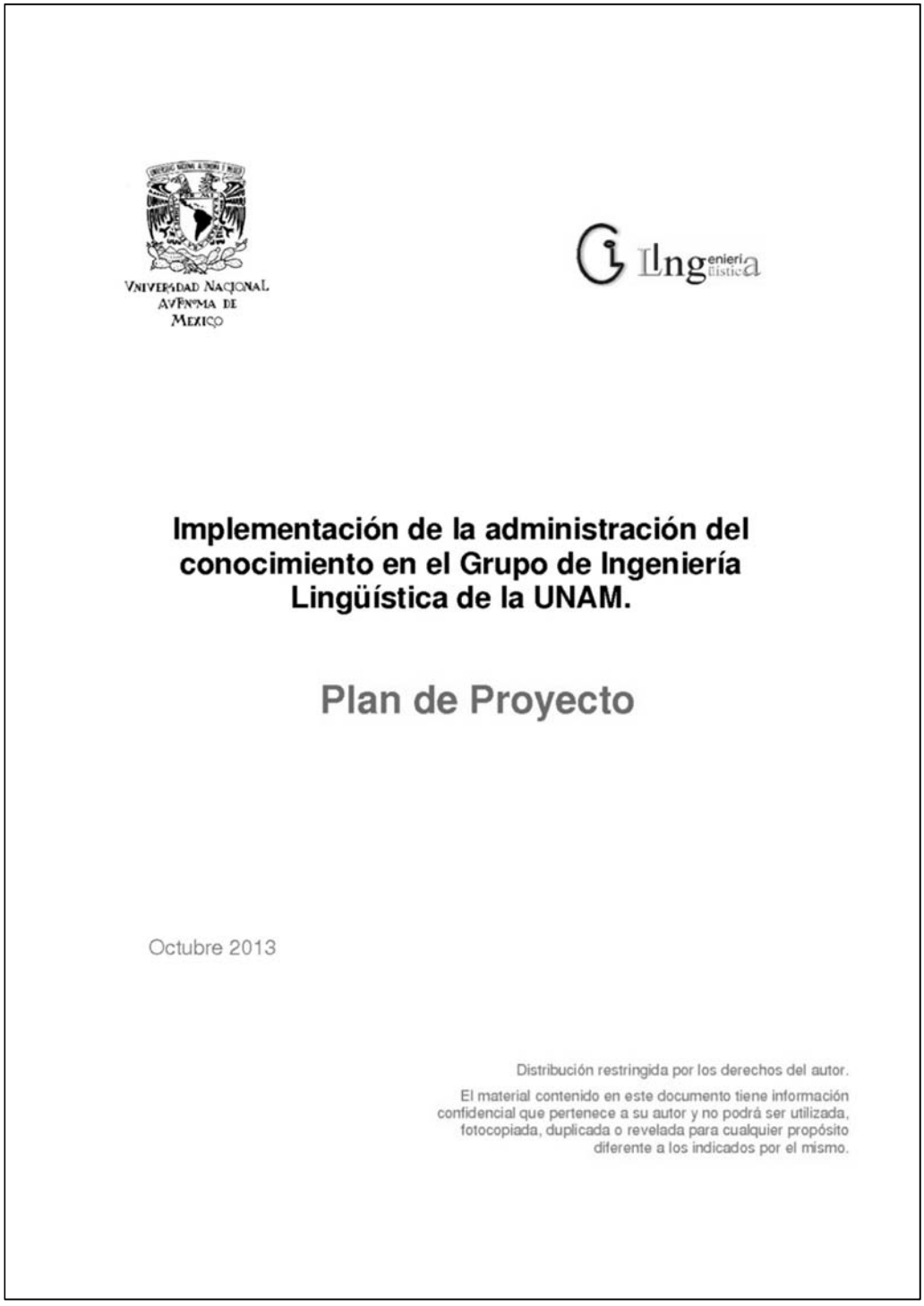
Como puede observarse, existen muchas opciones de trabajo futuro, tanto para mejorar la administración del conocimiento del Gil, como para optimizar la explotación y organización de su repositorio de conocimiento. Sin embargo, el propósito del presente trabajo de investigación queda concluido, pues se confirma la posibilidad de aplicar las tecnologías del lenguaje en la explotación y organización de repositorios de conocimiento.

El éxito en la utilización de dichas tecnologías del lenguaje en la administración del conocimiento de la institución utilizada como caso de estudio (el Gil), nos permite hacer una generalización, mediante un proceso de deducción y concluir que las tecnologías del lenguaje pueden aplicarse exitosamente en la administración del conocimiento para facilitar el desarrollo de sus actividades.

Anexos

Anexo 1

Plan de proyecto para la implementación de la administración del conocimiento en el Grupo de Ingeniería Lingüística de la UNAM. Se omiten algunas páginas por no tener relevancia.



VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

Ingeniería lística

Implementación de la administración del conocimiento en el Grupo de Ingeniería Lingüística de la UNAM.

Plan de Proyecto

Octubre 2013

1.1. INFORMACIÓN DEL DOCUMENTO

	Información
Identificador del Documento	Plan de Proyecto
Dueño del Documento	Grupo de ingeniería lingüística - UNAM
Autor del documento	Juan Luis Serralde Galicia
Fecha de Creación	10/oct/2013
Fecha de Última Modificación	22/oct/2013
Nombre del Archivo	KM_GIL_Plan_de_implementación.V.1.1

1.2. CONTROL DE CAMBIOS

Versión	Fecha del Cambio	Cambios
1.0	18/oct/2013	Documento Inicial
1.1	22/oct/2013	Corrección de la sección 2.18

1.3. REVISIONES

Revisores	Comentarios	Firmas	Fecha
Dr. Gerardo Sierra Dra. Azucena Móntez Mtro. Carlos Méndez	Cambiar el contenido de la sección 2.18		21/oct/2013

2.1. RESUMEN EJECUTIVO

El objetivo que persigue el presente proyecto es implementar el modelo de administración del conocimiento en el Grupo de Ingeniería Lingüística de la UNAM, en un lapso de 12 meses.

Esta implementación surge como respuesta a la necesidad de mejorar la eficiencia al compartir y acceder al conocimiento y la dependencia hacia individuos clave en la realización de actividades, por ser los únicos que poseen el conocimiento para realizarlas.
Al finalizar el proyecto el grupo trabajará inmerso en un modelo suficientemente maduro de administración del conocimiento que le permita identificar, documentar, almacenar, compartir y aplicar el conocimiento que genere.

A manera de valor agregado, el repositorio de información producido por la implementación de la administración del conocimiento servirá para hacer investigación (fuera del alcance de este proyecto) sobre extracción automática de conocimiento, usando técnicas y métodos de procesamiento textual.

2.2. INTRODUCCIÓN

Un elemento que comienza a enfocar la atención de las organizaciones como creador de valor es el conocimiento, más aún para un grupo dedicado a la generación del mismo, como es el caso del GIL-UNAM.

La alineación de la misión, visión y objetivos a la intención de generar conocimiento es común para este tipo de organizaciones, de ahí que se pre visualice necesaria su administración (al igual que cualquier tipo de capital) de manera que se facilite su creación, almacenamiento, acceso y aplicación. En respuesta a estas necesidades surge la Administración del conocimiento; esta pretende extraer el conocimiento implícito en la mente de las personas y aquel inmerso en documentos, para almacenarlos en un repositorio que facilite su acceso y extracción. También busca generar un esquema de trabajo que promueva la documentación constante del conocimiento generado y su continua aplicación por parte de los miembros de la organización.

Como se mencionó anteriormente, el conocimiento puede estar presente de manera explícita o tácita. En el primer caso, la información está plasmada en documentos, ya sea físicos o electrónicos, que pueden compartirse y ser recuperados fácilmente. Para este tipo de documentos, la extracción de información relevante y oportuna representa el problema más grande.

En el segundo caso, el conocimiento se encuentra de manera implícita, en la mente de los integrantes de la organización, a manera de experiencia, interpretación de las cosas y procesos de reflexión. En este tipo de conocimiento, el reto más grande radica en su extracción desde la mente del individuo.

2.3. VISIÓN DEL PROYECTO

En un año, el Grupo de Ingeniería Lingüística trabajará inmerso en un modelo de administración del conocimiento que le permita manejarlo como cualquier otro capital de la organización. Dicha administración permitirá identificar, documentar, almacenar, compartir y aplicar el conocimiento.

2.4. MISIÓN DEL PROYECTO

Implementar el modelo de administración del conocimiento en el Grupo de Ingeniería Lingüística de la UNAM tomando como referencia al APO KM Framework.

2.5. OBJETIVOS DEL PROYECTO

1. Preparar las bases culturales, técnicas y de estructura organizacional, necesarias para avalar una implementación exitosa de la Administración del Conocimiento.
2. Implementar el modelo de Administración del conocimiento
3. Revisar la correcta adopción del modelo implementado y realizar las posibles mejoras.

2.6. METAS DEL PROYECTO

#	Nombre de la meta	Descripción	Fecha de alcance
1	Preparación de las bases culturales, técnicas y de estructura organizacional.	Asegurar que existan las condiciones necesarias para poder realizar una implementación de la administración del conocimiento.	
1.1	Revisión de Misión, Visión y objetivos de la organización.	Revisión de los elementos mencionados y su posible actualización.	28/oct/2013
1.2	Obtención de recursos	Adquirir o asegurar la disponibilidad de los recursos tecnológicos, técnicos y humanos necesarios	30/oct/2013
1.3	Plan de cambio cultural	Creación de un plan para minimizar el rechazo al cambio por parte de los integrantes del GIL	8/nov/2013
2	Implementación de la administración del conocimiento	Realización de todas las actividades necesarias para poner en marcha el modelo de administración del conocimiento	
	Identificación del conocimiento a administrar	Identificar el tipo de conocimiento que se desea administrar, su origen y la forma de documentarlo.	13/nov/2013
2.1	Creación del repositorio	Implantar una herramienta web que permita el almacenamiento	27/nov/2013
2.2	Documentación del conocimiento	Plasmar en documentos el conocimiento que se desea administrar.	28/nov/2013 (inicio)
2.3	Almacenamiento del conocimiento	Ingresar en el repositorio los documentos creados.	28/mar/2013 (término)
2.4	Creación del nuevo modelo de trabajo en el GIL	Creación de un modelo de trabajo que garantice que el conocimiento sea documentado, almacenado, compartido y distribuido y aplicado.	10/dic/2013
3	Revisión de la implementación	Revisión del escenario presente para verificar el funcionamiento de la implementación realizada	26/may/2014
4	Análisis de oportunidades de mejora	Para este proyecto se limita el alcance a logar un nivel de madurez básico en la administración del conocimiento; sin embargo, esta actividad consiste en analizar posibles mejoras para un proyecto futuro. Este análisis servirá para documentar la	9/jun/2014

		experiencia adquirida y las particularidades observadas en la implementación.	
5	Entrega de documentación	La documentación de cada actividad se realiza a lo largo del proyecto. Al finalizar se integra a manera de memoria técnica para que el GIL tenga en su poder los detalles de la implementación, como configuraciones de servidores, contraseñas, formatos, documentos legales, etc.	23/jun/2014
6	Cierre del proyecto	Revisión del cumplimiento de metas, el transcurso del proyecto, el trabajo futuro, las recomendaciones y conclusiones. Al finalizar se realiza la firma del documento que exhibe el cierre del proyecto.	1/jul/2014

2.7. ALCANCES DEL PROYECTO

La administración del conocimiento implica un trabajo constante de documentar, compartir, almacenar y aplicar el conocimiento. Es un proceso cíclico e infinito, y en cada iteración se obtiene una madurez mayor del modelo. Sin embargo, el alcance de este proyecto se limita a una temporalidad de 12 meses, tiempo en el cual se espera tener un nivel funcional de administración del conocimiento que minimice el esfuerzo futuro para seguir mejorando.

En la actividad de "identificación" se encuentran las áreas de conocimiento involucradas en la organización y se jerarquizan en relación a la importancia que estas tienen en la cadena productiva de valor de la organización. En una administración del conocimiento con un nivel total de madurez se administran todas las áreas; sin embargo, parta el presente trabajo solo se administrarás las 2 más altas en valor jerárquico.

2.8. PROBLEMÁTICA QUE RESUELVE

- Limitada administración del capital intelectual del GIL.
- Poca eficiencia al compartir conocimiento, producto del almacenamiento no centralizado.
- Curva de aprendizaje alta en relación al conocimiento de los procesos internos (administrativos y operativos) y la identidad de la organización.
- Dependencia hacia individuos clave en la realización de actividades por ser los únicos que poseen el conocimiento para realizarlas.

2.9. BENEFICIOS ESPERADOS Y PRINCIPALES BENEFICIARIOS

Beneficio	Beneficiarios
Fácil y rápido acceso al conocimiento de la organización desde un repositorio centralizado.	Integrantes del GIL
Extracción del conocimiento implícito en la mente de los integrantes del GIL y su almacenamiento para poder usarlo como un capital de la organización.	Integrantes del GIL, Directivos del GIL, Sociedad
Disminución de la curva de aprendizaje	Integrantes del GIL, Directivos del GIL
Facilidad para compartir y aplicar el conocimiento	Integrantes del GIL
Agilización para una mejora o reingeniería de los procesos internos	Integrantes del GIL, Directivos del GIL
Mejora y reingeniería de los procesos internos.	Integrantes del GIL, Directivos del GIL

2.10. ESTRUCTURA ORGANIZACIONAL DEL EQUIPO DEL PROYECTO

2.11. PRINCIPALES INVOLUCRADOS

Líder de proyecto
Responsable: Lic. I. Juan Luis Serralde Galicia
Responsabilidades:

- Administrar la integración del proyecto.
- Crear y dar seguimiento del plan de trabajo.
- Controlar el progreso del proyecto. Alcance de metas en tiempo y forma.
- Administrar los recursos asignados.
- Divulgar proactivamente la información a los involucrados en el proyecto.
- Manejar los riesgos identificados.
- Realizar las entregas esperadas y hacer el cierre al término del proyecto.

Dirección del Grupo de Ingeniería Lingüística, UNAM
Responsable: Dr. Gerardo Sierra
Responsabilidades:

- Aprobar el plan de proyecto.
- Revisión de avances.
- Firmar los documentos necesarios para permitir el desarrollo ágil del proyecto.
- Proveer los requerimientos técnicos, tecnológicos y humanos necesarios.
- Garantizar el apoyo del GIL en la realización de las tareas que se les encomienden.
- Apoyar las actividades descritas en el Plan de cambio cultural.
- Cierre del proyecto.

Desarrollador web
Responsabilidades:

- Crear el repositorio del conocimiento.
- Administrar el repositorio y los recursos tecnológicos asignados.

Documentador
Responsabilidades:

- Documentar los procesos de la organización usando como referencia el modelo de ITIL.

2.12. TIEMPO ESTIMADO

12 meses

2.13. PLAN DE ACTIVIDADES

2.14. MATRICES DE RESPONSABILIDADES

Metas	
Clave	Descripción
A1	Preparación de las bases culturales, técnicas y de estructura organizacional.
A2	Revisión de Misión, Visión y objetivos de la organización.
A3	Obtención de recursos.
A4	Plan de cambio cultural.
A5	Implementación de la administración del conocimiento.
A6	Identificación del conocimiento a administrar.
A7	Creación del repositorio.
A8	Documentación del conocimiento.
A9	Almacenamiento del conocimiento.
A10	Creación del nuevo modelo de trabajo en el GIL.
A11	Revisión de la implementación.
A12	Análisis de oportunidades de mejora.
A13	Entrega de documentación.
A14	Cierre del proyecto.

Responsables	
Clave	Nombre
R1	Lic. I. Juan Luis Serralde Galicia
R2	Dr. Gerardo Sierra
R3	Desarrollador web
R4	Documentador

	R1	R2	R3	R4
A1	A, R, C, I	C, I		
A2	R, C, I	A, R, C, I		
A3	A	R		
A4	R, C, I	A, R, C, I		
A5	A, R, C	I		
A6	A, R, C, I	C, I		
A7	A, C		R	
A8	A	C		R
A9	A			R
A10	R, C	A, R		
A11	A, R			
A12	A, R	I		
A13	A, R	I		
A14	A, R	R		

Notación:

A: Accountable
R: Responsible
C: Consulted
I: Informed

2.15. FACTORES CRÍTICOS DE ÉXITO

- Contar con el apoyo de los directivos del GIL para la realización de las actividades descritas en este plan de proyecto.
- Disponibilidad de los recursos indicados en el presente documento.
- Adopción por parte de todos los integrantes del GIL del nuevo modelo de trabajo planteado en la implementación de la administración del conocimiento.

2.16. RIESGOS

Riesgo	Probabilidad	Impacto	Calificación	Controles
Rechazo al cambio.	Alta	Alto	Alto	1, 2
Disminución paulatina del seguimiento del avance del proyecto por parte del líder del proyecto y de los directivos del GIL	Baja	Alto	Medio	3
Para el caso del desarrollador web o documentador: Abandono de las actividades asignadas.	Media	Medio	Medio	4
Bajo nivel de conocimiento de parte del desarrollador del repositorio de conocimiento.	Baja	Medio	Baja	4

Actividades no contempladas en el plan de proyecto, pero necesarias para la implementación	Alto	Medio	Alto	5
Falta de conocimiento profundo sobre implementación del modelo de administración del conocimiento por parte del líder del proyecto	Medio	Alto	Alto	5
Pérdida de interés de los directivos del GIL	Baja	Alto	Alto	3

Controles			
#	Nombre del control	Descripción	Tipo
1	Creación del plan de cambio cultural	Elaboración de un plan de actividades que minimicen el impacto del cambio	Prevenir, Mitigar
2	Exhibición del interés de los directivos del GIL	Buscar medios para exponer el interés de los directivos del GIL en aplicar la administración del conocimiento	Prevenir. Mitigar
3	Calendarización de juntas	Programación de citas semanales para revisar los avances en la implementación	Prevenir
4	Selección de perfiles adecuados	Realizar un análisis de los perfiles requeridos para un correcta realización de las actividades. Se contemplan tiempos disponibles, conocimientos, capacidades, planes de vida y carrera.	Prevenir
5	Revisión de proyectos similares	Análisis de la literatura existente sobre la implementación de la administración del conocimiento en otras organizaciones.	Prevenir Mitigar, Restaurar

2.17. DESCRIPCIÓN DE ENTREGABLES

Nombre del entregable	Descripción	Fecha de entrega
Plan de proyecto	Contiene el plan de proyecto de implementación de KM en el GIL-UNAM	18/oct/2013
Documentación técnica y administrativa	Manuales, planes de trabajo, diagramas, memorias técnicas y demás elementos que documentan la información propia del proyecto.	17/oct/2014
Acta de cierre de proyecto	Documento que registra la finalización del proyecto.	18/oct/2014

2.18. CONDICIONES, EXCEPCIONES Y RESTRICCIONES

- Es indispensable el apoyo continuo de los administrativos del GIL para garantizar la adopción del modelo implementado.
- El desarrollo de este proyecto se limita a un año, por lo que es esencial dar un seguimiento constante a fin de asegurar el logro de metas en el tiempo planeado.
- Considerando la limitante de tiempo, es necesario contar con la aprobación del plan de proyecto y la asignación de los recursos descritos para iniciar la implementación con la mayor anticipación posible.

2.19. RECURSOS

#	Descripción	Costo mensual
1	Líder de proyecto (Implementador)	$ 15,000
2	Desarrollador web	$ 10,000
3	Documentador	$ 5,000 *
4	PC para cada integrante del equipo de implementación (1-3)	$ 50 *
6	Papelería	$ 100 *
9	Espacios de trabajo para el equipo de desarrollo (1-3)	$ 2000 *
12	Software para la creación del repositorio	$ 0

Nota: Los montos corresponden a los costos estimados y sirven para analizar la pertinencia del proyecto. El Grupo de Ingeniería Lingüística ya cuenta con los recursos listados anteriormente, de manera que no será necesario hacer ninguna erogación extra.

* Costos calculados por la técnica del prorrateo.

2.20. ANÁLISIS COSTO-BENEFICIO

El Grupo de Ingeniería Lingüística cuenta con los recursos humanos, técnicos y tecnológicos suficientes para la realización de este proyecto, de manera que los costos incurridos por su realización son nulos. En contraste, los beneficios esperados ya descritos en una sección anterior, justifican el tiempo y esfuerzo requerido para la implementación.
En resumen, es un proyecto viable, oportuno y valioso.

2.21. CONCLUSIONES

La realización del presente proyecto se manifiesta viable y necesaria para generar mejoras en la forma de crear, almacenar, compartir y aplicar el conocimiento surgido en el Grupo de Ingeniería Lingüística de la UNAM. A un año de haber comenzado este proyecto, el grupo trabajará inmerso en un modelo razonablemente maduro de administración del conocimiento que le genere las capacidades que promete en la teoría.

El éxito en la implementación será consecuencia de la participación activa de todos los miembros del GIL, una correcta planeación y el impetuoso y constante esfuerzo del equipo de desarrollo.
Finalmente, la experiencia que emane de la implementación tendrá el valor de ser parte del conocimiento pionero en materia de implementación de administración del conocimiento en un grupo de investigación.

Anexo 2

Carta aceptación del plan de proyecto para la implementación de la administración del conocimiento en el Grupo de Ingeniería Lingüística de la UNAM.

	Proyecto de implementación de KM en el GIL-UNAM	

ACTA DE ACEPTACIÓN DEL PLAN DE PROYECTO

Los abajo firmantes hemos leído y analizado el **plan de proyecto para la Implementación de un modelo de administración del conocimiento en el Grupo de Ingeniería Lingüística de la UNAM**, plasmado en el documento llamado **KM_GIL_Plan_de_implementación.V.1.1**, y damos el visto bueno para la realización del proyecto en los términos ahí descritos. Al mismo tiempo, asumimos una posición de compromiso y apoyo al proyecto y al equipo de trabajo para lograr los objetivos planteados.

Rol	Nombre	Firma
Jefe del GIL (Grupo de Ingeniería Lingüística)	Dr. Gerardo Eugenio Sierra Martínez	
Representante del GIL	Dra. Azucena Montes Rendón	
Representante del GIL	Dr. Carlos Francisco Méndez Cruz	
Líder de proyecto	Lic. Juan Luis Serralde Galicia	

Ciudad Universitaria, México D.F. a 25 de Octubre de 2013

Anexo 3

APO KM Assessment Tool (Traducido al español)

Evaluación inicial

INFORMACIÓN RELEVANTE

El Grupo de Ingeniería Lingüística (GIL) de la UNAM pretende implementar un modelo de administración del conocimiento que le permita mejorar la eficiencia al compartir y acceder al conocimiento, disminuir la dependencia hacia individuos clave en la realización de actividades (por ser los únicos que poseen el conocimiento para realizarlas) y acortar la curva de aprendizaje de los nuevos integrantes en relación al conocimiento de los procesos internos (administrativos y operativos) y la identidad de la organización.

Para tal efecto, se ha desarrollado un plan de proyecto que define, entre otras actividades, la elaboración de una valoración previa a la implementación del modelo, con la finalidad de tener referencias que permitan comparar los logros a determinado tiempo. El objetivo de este cuestionario es realizar dicha valoración inicial del Nivel de Madurez de la Administración del Conocimiento en el GIL.

Le invitamos a que nos ayude a realizar esta valoración contestando el presente cuestionario y, de antemano, le agradecemos el tiempo que le dedique a esta tarea que está diseñada para realizarse en un máximo de 10 minutos.

Aprovechamos para manifestar por escrito que los datos y las respuestas que nos entregue serán confidenciales y únicamente serán analizados y procesados por el implementador de la administración del conocimiento: Lic. I. Juan Luis Serralde. El implementador está comprometido a no divulgar la información contenida y a hacer uso de ellos única y exclusivamente para los fines aquí descritos.

El cuestionario no contempla respuestas correctas o incorrectas, siéntase en la libertad de contestar lo que a su parecer refleje en mayor grado su visión sobre el fenómeno.

INSTRUCCIONES

Para fines de este documento se hará uso de la palabra "organización" para referirnos al GIL-UNAM y se entenderá por "empleados" a todos sus integrantes.

El presente cuestionario consta de 42 reactivos divididos en 7 categorías. Cada categoría está presentada en forma de una tabla que contiene 6 frases. Para cada Frase (ítem) colocada al lado izquierdo de la tabla existe un espacio del lado derecho que servirá para que usted coloque el valor que considere pertinente para evaluar el grado en que se realizan las acciones indicadas en la frase. La escala que se utilizará será la siguiente:

Descripción	Escala de calificación
Haciéndolo muy bien	5
Haciéndolo bien	4
Haciéndolo adecuadamente	3
Haciéndolo limitadamente	2
Haciéndolo muy limitadamente o no se hace nada	1

Al final de cada categoría encontrará una fila con una frase similar a "Subtotal cat. x", Por favor, omita esa línea y continúe con la siguiente tabla (categoría).
Al terminar entregue el cuestionario al aplicador.

CUESTIONARIO

Información del participante	
Nombre	
Área	◯Operativa ◯Directiva
Sexo	◯Masculino ◯Femenino
Rango de edad (años)	◯Menos de 20 ◯21-30 ◯31-40 ◯41-50 ◯Más de 50
Nivel académico (Estudios terminados)	◯Carrera técnica ◯Licenciatura ◯Maestría ◯Doctorado
Antigüedad en la organización	◯Menos de 6 meses ◯6 meses o más

	cat 1.0. Liderazgo	Calificación
1	En la organización se comparte una misma visión del conocimiento y se tiene una estrategia fuertemente ligada a la visión, misión y objetivos de la organización.	
2	La organización ha tomado medidas para formalizar las iniciativas de KM (que son: la creación de la unidad central de coordinación para la KM, la designación de jefe de conocimiento y oficial de la información y la creación de redes de conocimiento).	
3	Hay fondos que se destinan a las iniciativas de KM.	
4	La organización cuenta con una política para salvaguardar el conocimiento (es decir, derechos de autor, patentes, administración de conocimiento y la política de seguridad del conocimiento).	
5	Los gerentes son un modelo a seguir de los valores de intercambio del conocimiento y del trabajo colaborativo. Los gerentes invierten más tiempo en la difusión de la información a su personal y facilitando el flujo de información de su personal con otros departamentos, divisiones o unidades.	
6	La Dirección promueve, reconoce y premia las mejoras del desempeño, el aprendizaje organizacional e individual, el intercambio de conocimiento y la creación de conocimiento y la innovación.	
	Subtotal cat 1.0: Liderazgo de la Administración del conocimiento	

	cat 2.0. Procesos	Calificación
7	La organización determina sus principales competencias (capacidades estratégicas importantes que proporcionan una ventaja competitiva) y las alinea a su misión y objetivos estratégicos.	
8	La organización diseña su sistema de trabajo y sus procesos clave para crear valor a los clientes y lograr la excelencia en su desempeño.	
9	La nueva tecnología, el intercambio del conocimiento en la organización, la flexibilidad, la eficiencia y la eficacia son factores considerados en el diseño de procesos.	
10	La organización cuenta con un sistema organizado para la administración de situaciones de crisis o eventos imprevistos que garantiza una operación ininterrumpida, la prevención y recuperación.	
11	La organización implementa y administra sus procesos de trabajo clave para asegurar que los requisitos del cliente se cumplan y los resultados del negocio se logren.	
12	La organización evalúa y mejora continuamente sus procesos de trabajo para lograr un mejor desempeño, para disminuir variaciones, mejorar los productos y servicios, y para estar actualizados con las últimas tendencias, desarrollos y direcciones de negocio.	
	Subtotal cat 2.0: Procesos	

	cat 3.0. Gente	Calificación
13	La educación, el entrenamiento y el programa de desarrollo de carrera de la organización permiten desarrollar los conocimientos, habilidades y capacidades de los empleados, los apoya al logro de los objetivos y contribuye a generar un alto desempeño.	
14	La organización cuenta con un proceso sistemático de inducción para el nuevo personal que incluye la familiarización con la Administración del Conocimiento y sus beneficios, el sistema de Administración del Conocimiento y sus herramientas.	
15	La organización cuenta con procesos formales de consejo y guía (mentoring), asesoría (coaching) y tutoría.	
16	La organización cuenta con una base de datos de las competencias del personal.	
17	El intercambio de conocimiento y la colaboración son activamente impulsados y premiados o corregidos.	
18	Los empleados están organizados en pequeños equipos (por ejemplo, círculos de calidad, equipos de mejora de trabajo, equipos multidisciplinarios, comunidades de práctica) para responder a los problemas del ámbito laboral.	
	Subtotal cat 3.0: Gente	

	cat 4.0. Tecnología	Calificación
19	La dirección ha establecido una infraestructura de TI (es decir, internet, intranet y página web) y se han desarrollado capacidades para facilitar una Administración de Conocimiento efectiva.	
20	La infraestructura de TI está alineada con la estrategia de Administración de Conocimiento de la organización.	
21	Todo el personal tiene acceso a una computadora.	
22	Todo el personal tiene acceso a internet / intranet y a una dirección de correo electrónico.	
23	La información puesta a disposición en el sitio web / intranet se actualiza de manera regular.	
24	La intranet (o red similar) es utilizada como la fuente principal de comunicación de toda la organización para apoyar la transferencia del conocimiento o el intercambio de información.	
	Subtotal cat 4.0: Tecnología	

	cat 5.0. Procesos de Administración del Conocimiento	Calificación
25	La organización cuenta con procesos sistemáticos para identificar, crear, almacenar, compartir y aplicar conocimiento.	
26	La organización mantiene un inventario de conocimiento en el que se identifican y ubican los activos de conocimiento o recursos de toda la organización.	
27	El conocimiento que se recopila de las tareas o proyectos terminados está documentado y se comparte.	
28	Se retiene el conocimiento crítico de los empleados que dejan la organización.	
29	La organización comparte las mejores prácticas y las lecciones aprendidas en toda la organización para que no se vuelva a reinventar la rueda y se duplique el trabajo.	
30	Las actividades para el estudio comparativo de mercado se realizan con enfoque interno y externo a la organización, y sus resultados son utilizados para mejorar el desempeño organizacional y para crear nuevo conocimiento.	
	Subtotal cat 5.0: Procesos de conocimiento	

	cat 6.0. Aprendizaje e Innovación	Calificación
31	La organización comunica y refuerza continuamente los valores de aprendizaje e innovación.	
32	La organización considera tomar riesgos o cometer errores como oportunidades de aprendizaje siempre y cuando no ocurran de manera repetida.	
33	Los equipos multidisciplinarios se organizan para hacer frente a los problemas que atraviesan diferentes unidades de la organización.	
34	El personal siente que está empoderado y que sus ideas y contribuciones generalmente son valoradas por la organización.	
35	La dirección está dispuesta a probar nuevas herramientas y métodos.	
36	Se incentiva al personal para que trabaje en conjunto y comparta información.	
	Subtotal cat 6.0: Aprendizaje de innovación	

	cat 7.0. Resultado de la Administración del Conocimiento	Calificación
37	La organización cuenta con un historial de implementación exitosa de la Administración del conocimiento y de otras iniciativas de cambio.	
38	Existen medidas para evaluar el impacto de las contribuciones e iniciativas de conocimiento.	
39	La organización ha logrado una mayor productividad través de la reducción del ciclo de tiempo, grandes ahorros de costos, mejora de la efectividad, uso más eficiente de los recursos (incluyendo el conocimiento), mejora en la toma de decisiones y un incremento en la velocidad de la innovación.	
40	La organización ha aumentado su rentabilidad como resultado de la productividad, la calidad y la mejora de la satisfacción del cliente.	
41	La organización ha mejorado la calidad de sus productos y / o servicios como resultado de la aplicación del conocimiento para mejorar los procesos de negocios o las relaciones con los clientes.	
42	La organización ha crecido de manera sostenida como resultado de mayor productividad, mayor rentabilidad y mejor calidad de los productos y los servicios.	
	Subtotal cat 7.0: Resultado de la Administración del Conocimiento	

Anexo 4

Misión, visión y objetivos del Grupo de Ingeniería Lingüística de la UNAM

Misión:

Generar conocimiento básico y aplicado en el área de la ingeniería lingüística.

Visión:

Ser el grupo líder en investigación y desarrollo tecnológico del área de la ingeniería lingüística.

Objetivos:

1. Formación de recursos humanos en el área de la ingeniería lingüística
   1.1 Pertenecer a programas de posgrado
   1.2 Fomentar la participación de los miembros del GIL en los programas académicos de la UNAM

2 Difundir el área de la ingeniería lingüística
   2.1 Participación en congresos nacionales e internacionales
   2.2 Organización del seminario de ingeniería lingüística
   2.3 Participación en foros empresariales
   2.4 Utilización de recursos web para dar a conocer la labor del GIL.

3 Desarrollar sistemas informáticos, desarrollos tecnológicos, herramientas, tecnologías, métodos y productos que permitan resolver problemas sociales.
   3.1 Adoptar prácticas de administración de proyectos
   3.2 Tener un área de desarrollo de productos

4 Realizar investigación en el área de ingeniería lingüística
   4.1 Escribir artículos científicos y libros
   4.2 Mantener un seminario interno de discusión
   4.3 Vinculación con instituciones de nivel superior
   4.4 Adquisición de acervo bibliográfico
   4.5 Dirección de tesis
   4.6 Participación en congresos
   4.7 Evaluación constante de líneas de investigación

5 Tener proyectos de investigación en el área de ingeniería lingüística
   5.1 Participación en proyectos de CONACYT
   5.2 Participación en foros empresariales para la atracción de proyectos patrocinados.

Anexo 5

Taxonomía del conocimiento del GIL

Dimensión (Nivel 1)	Conocimiento del GIL	
Descripción de la dimensión	Contiene todo el conocimiento que el GIL posee para llevar a cabo sus actividades cotidianas	
Nivel 2	**Nivel 3**	**Comentarios**
Investigación		Contiene el conocimiento que el GIL posee referente a sus tareas de investigación
	Acervo Bibliográfico	Contiene el acervo bibliográfico digital con el que cuenta el GIL
	Corpus	Contiene los corpus (Conjuntos estructurados de textos) utilizados en las investigaciones del GIL
	Herramientas	Contiene las herramientas de software que se ocupan comúnmente en las tereas de investigación del GIL
	Publicaciones GIL	Publicaciones realizadas por los integrantes del GIL o donde estos participaron como colaboradores
	Proyectos	Contiene la documentación de los proyectos de investigación que ha realizado o realiza actualmente el GIL
Formación de Recursos Humanos		Conocimiento que el GIL posee en relación a su formación de Recursos Humanos
	Cursos	Contiene los cursos en video creados por los integrantes del GIL
	Publicaciones GIL	Publicaciones realizadas por los integrantes del GIL o donde estos participaron como colaboradores
	Acervo bibliográfico	Contiene el acervo bibliográfico digital con el que cuenta el GIL
Difusión		Conocimiento que el GIL posee en relación a sus procesos de difusión
	Publicaciones GIL	Publicaciones realizadas por los integrantes del GIL o donde estos participaron como colaboradores
	Eventos	Contiene la documentación de los eventos de difusión realizados y los que están por realizarse
Desarrollo de sistemas y tecnologías		Conocimiento que el GIL posee para desarrollar sistemas y tecnologías
	Sistemas de consulta de corpus	Contiene los accesos a los sistemas de consulta de corpus y su documentación
	Recursos para el desarrollo de sistemas	Conocimiento y herramientas que el GIL ocupa comúnmente para desarrollar sistemas

	Sistemas de extracción conceptual	Contiene los accesos a los sistemas de extracción conceptual y su documentación
	Minería de textos	Contiene los accesos a los sistemas de minería de textos y su documentación
	Diccionarios electrónicos	Contiene los accesos a los diccionarios electrónicos que usa cotidianamente el GIL
	Sistemas de análisis forense	Contiene los accesos a los sistemas de consulta de corpus y su documentación
Administración del laboratorio		Conocimiento que el GIL posee para poder administrar su laboratorio
	Directorio	Contiene la relación de nombres, puestos y teléfonos del personal que el GIL contacta frecuentemente para realizar sus procesos cotidianos
	Asignación de recursos	Contiene el conocimiento relacionado con la administración de los recursos materiales, tecnológicos y de espacios del laboratorio del GIL
	Expedientes	Contiene los expedientes de los integrantes del GIL
	Procedimientos	Contiene la documentación de los procesos administrativos del GIL
	Galería	Conjunto de imágenes de los eventos del GIL
	Recursos para la administración	Recursos que son usados cotidianamente en los procesos administrativos del GIL. Por ejemplo, logotipos oficiales, formatos, hojas membretadas, etc.
	Cómputo	Documentación relacionada a la configuración de la infraestructura de cómputo del GIL
Vinculación exterior		Conocimiento que el GIL posee para realizar sus actividades de vinculación exterior
	Exentos	Contiene la documentación de los eventos de vinculación realizados y los que están por realizarse

Anexo 6

Conjunto de experimentos realizados en el ejercicio de clustering

#	Tamaño del corpus	Preprocesamiento del texto	Tipo de matriz	Algoritmo	Parámetros del algoritmo	Clusters generados	Resultado	Notas
1	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas	Simple kMeans	Número de clusters: 2 Distancia: euclidiana Semilla: Manual (6) Split: 70% (entrenamiento) 30% (Prueba)	Cluster 0: 0 u Cluster 1: 13 u Total: 13 unidades (30%)	Todas las unidades de prueba las asigna al cluster 1	Se cambió la semilla por los valores 12, 18, 20 y el resultado fue el mismo
2	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas	Simple kMeans	Número de clusters: 3 Distancia: euclidiana Semilla: Manual (6) Split: 50% (entrenamiento) 50% (Prueba)	Cluster 0: 0 u Cluster 1: 20 u Cluster 2: 1 u Total: 21 unidades (50%)	Casi todas las unidades de prueba las asigna al cluster 1	
3	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas	Simple kMeans	Número de clusters: 5 Distancia: euclidiana Semilla: Manual (18) Split: 50% (entrenamiento) 50% (Prueba)	Cluster 0: 0 u Cluster 1: 20 u Cluster 2: 1 u Cluster 3: 0 u Cluster 4: 0 u Total: 21 unidades (50%)	Casi todas las unidades de prueba las asigna al cluster 1	
4	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas	EM	Número de clusters: automático Semilla: Manual (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 E-6	Cluster 0: 1 u Cluster 2: 7 u Cluster 3: 2 u Cluster 4: 1 u Cluster 5: 1 u Cluster 7: 9 u Total: 21 unidades (50%)	Ya se comienzan a distribuir las unidades. Al parecer existen 3 clusters	
5	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas con stop words	EM	Número de clusters: 3 Semilla: Manual (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 E-6	Cluster 0: 0 u Cluster 1: 18 u Cluster 2: 3 u Total: 21 unidades (50%)	El experimento muestra solo 2 clusters	Se obtiene el mismo resultado si se considera la mínima desviación estándar como E-4 o E-2
6	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas con stop words	EM	Número de clusters: 5 Semilla: Manual (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 E-6	Cluster 0: 9 u Cluster 1: 11 u Cluster 2: 1 u Total: 21 unidades (50%)	El experimento muestra solo 2 clusters principales	
7	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas con stop words	EM	Número de clusters: 5 Semilla: Manual (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 **E-4**	Cluster 0: 9 u Cluster 1: 11 u Cluster 2: 1 u Total: 21 unidades (50%)	Se repite completamente el mismo patrón del experimento anterior	
7.1	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias absolutas con stop words	EM	Número de clusters: Automático Semilla: Manual (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 **E-6**	Cluster 0: 1 u Cluster 1: 7 u Cluster 2: 2 u Cluster 3: 1 u Cluster 4: 1 u Cluster 5: 9 u Total: 21 unidades (50%)		
8	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias relativas	Simple kMeans	Número de clusters: 3 Distancia: euclidiana Semilla: Manual (6) Split: 50% (entrenamiento) 50% (Prueba)	Cluster 0: 0 u Cluster 1: 20 u Cluster 2: 1 u Total: 21 unidades (50%)	casi todas las unidades de prueba las asigna al cluster 1	Se obtiene el mismo resultado que el experimento 2 con la misma configuración del algoritmo pero ahora usando frecuencias relativas sin stop words

9	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias relativas	EM	Número de clusters: 3 Distancia: euclidiana Semilla: Manual (6) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 E-6	Cluster 0: 0 u Cluster 1: 18 u Cluster 2: 3 u Total: 21 unidades (50%)	Aparentemente solo existen 2 clusters	
10	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias relativas	EM	Número de clusters: 3 Distancia: euclidiana Semilla: Manual (6) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:1.0 E-2	Cluster 0: 0 u Cluster 1: 21 u Cluster 2: 03 u Total: 21 unidades (50%)	Todos los manda a un solo cluster	Si disminuimos la MinStdDev permitimos que un cluster admita más elementos distantes así que el número de clusters encontrados disminuye
11	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias relativas	EM	Número de clusters: automático Distancia: euclidiana Semilla: Manual (6) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-12**	Cluster 0: 10 u Cluster 1: 1 u Cluster 2: 7 u Cluster 3: 3 u Total: 21 unidades (50%)	El algoritmo sigue sugiriendo la existencia de 3 clusters principales	Se observólo siguiente Cluster 0: 10 u (ligüística) Cluster 1: 1 u (outlier) Cluster 2: 7 u (Computación) Cluster 3: 3 u (Doctorado)
12	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias relativas con stop words	EM	Número de clusters: automático Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-12**	Cluster 0: 9 u Cluster 1: 1 u Cluster 2: 8 u Cluster 3: 3 u Total: 21 unidades (50%)	El algoritmo sigue sugiriendo la existencia de 3 clusters principales	Se observó lo siguiente Cluster 1: 1 u (outlier, ingeniero civil) Fernando Arancibia Los otros clusters son difícil de diferenciar
13	41 tesis	Se eliminan caracteres no alfanuméricos	Frecuencias relativas con stop words	EM	Número de clusters: 3 Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-12**	Cluster 0: 11 u Cluster 1: 9 u Cluster 2: 1 u Total: 21 unidades (50%)		
14	40 tesis (Se eliminó el outlier)	Se eliminan caracteres no alfanuméricos	Frecuencias relativas con stop words	EM	Número de clusters: automático Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-12**	Cluster 0: 10 u Cluster 1: 7 u Cluster 2: 3 u Total: 20 unidades (50%)		Al eliminar el oulier se esperaba eliminar el cluster que lo contenía ya que era el único en ese cluster. El resultado demostró lo esperado.
15	40 tesis (Se eliminó el outlier)	Se eliminan caracteres no alfanuméricos	TF-IDF	EM	Número de clusters: automático Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-12**	Cluster 0: 20 u Total: 20 unidades (50%)	Encuentra un único cluster	Hay un documento muy pequeño (3KB) que solo contiene la carátula y el índice. Cuando se calcula el tfidf, las palabras comunes a todos los documentos no se eliminan (tfidf=0) porque no aparecen en ese documento pequeño. Con esto, pareciera que todos los demás documentos pertenecen a un mismo y único cluster.
16	38 tesis (Se eliminó el outlier)	Se eliminan caracteres no alfanuméricos	TF-IDF	EM	Número de clusters: automático Semilla: Default (100)	Cluster 0: 14 u Cluster 1: 4 u Cluster 2: 1 u		Para este experimento se eliminó el documento que pesa menos de 3Kb.

	Se eliminó el archivo que pesa menos de 3 KB Se eliminó un archivo que tenía nombre distinto pero era el mismo				Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-6**	Total: 19 unidades (50%)		También se eliminó un archivo repetido, con diferente nombre pero con contenido igual. Se calcularon nuevamente los valores de tfidf
17	38 tesis (Se eliminó el outlier) Se eliminó el archivo que pesa menos de 3 KB Se eliminó un archivo que tenía nombre distinto pero era el mismo	Se eliminan caracteres no alfanuméricos	TF-IDF	EM	Número de clusters: automático Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-12**	Cluster 0: 19 u Total: 20 unidades (50%)	Encuentra un único cluster	Cuando indicamos una desviación estándar de 1.0 E-12 , se forma un púnico cluster. Al parecer no era necesario recalcular tfidf y se intuye que el arcivo peroeño (3KB) no afectaba al ejercicio de clustering
18	38 tesis (Se eliminó el outlier) Se eliminó el archivo que pesa menos de 3 KB Se eliminó un archivo que tenía nombre distinto pero era el mismo	Se eliminan caracteres no alfanuméricos	TF-IDF	EM	Número de clusters: automático Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-2**	Cluster 0: 1 u Cluster 2: 3 u Cluster 3: 15 u Total: 19 unidades (50%)		
19	38 tesis	Se eliminan caracteres no alfanuméricos	TF-IDF	Simple kMeans	Número de clusters: 3 Distancia: euclidiana Semilla: Manual (10) Split: 50% (entrenamiento) 50% (Prueba)	Cluster 1: 19 u	Encuentra un único cluster	kMeans da casi el mismo resultado que con frecuencias absolutas. Cuando el algoritmo usa valores tfidf identifica a todos los elementos como similares y los asocia a un solo cluster.
20	38 tesis	Se eliminan caracteres no alfanuméricos	TF-IDF	Simple kMeans	Número de clusters: 3 Distancia: Manhattan Semilla: Manual (10) Split: 50% (entrenamiento) 50% (Prueba)	Cluster 1: 19 u	Encuentra un único cluster	Cambiando manualmente la semilla se obtiene el mismo resultado

21	38 tesis	Se eliminan caracteres no alfanuméricos	TF-IDF sin columnas completas con ceros	EM	Número de clusters: automático Semilla: Default (100) Split: 50% (entrenamiento) 50% (Prueba) MinStdDev:**1.0 E-6**	Cluster 0: 1 u Cluster 2: 3 u Cluster 3: 15 u Total: 19 unidades (50%)		

Anexo 7

Clusters formados por algunos experimentos de clustering.

No. Experimento: 7.1		
0	1	2
Ejuliovargasmejia_2013.txt (L-Inf) OPTIMIZACIÓN DEL GENERADOR DE CONCORDANCIAS DEL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO CHEM MEDIANTE UNA COMPARACIÓN ENTRE MANEJADORES DE BASES DE DATOS RELACIONALES Y ADMINISTRACIÓN DE DESEMPEÑO DE BASES DE DATOS Ejorgelazarohernandez_2010.txt (L-lylh) EXTRACCIÓN DE LA TERMINOLOGÍA BÁSICA DE LAS SEXUALIDADES EN MÉXICO A PARTIR DE UN CORPUS LINGÜÍSTICO Elauraelenahernandez.txt (M-cienccomp) CREACIÓN SEMI AUTOMÁTICA DE LA BASE DE DATOS Y MEJORA DEL MOTOR DE BÚSQUEDA DE UN DICCIONARIO ONOMASIOLÓGICO Ealejandrorosas_2011.txt (L-lylh) ANÁLISIS ESTILOMÉTRICO PARA LA DETECCIÓN DE PLAGIO Eantoniaestradatrejo_2009.txt (L-lylh) ANÁLISIS DE LAS RELACIONES CONCEPTUALES EN UNA TERMINOLOGÍA APLICACIÓN AL DICCIONARIO DE LINGÜÍSTICA Ecesarantonioaguilar_2008.txt (D-Ling) Análisis lingüístico de definiciones en contextos Definitorios Evictorsanchez-marcovelazquez_2013.txt (L-ingcomp) DESARROLLO DE UN EXTRACTOR DE PALABRAS CLAVE Eirasemacruzdominguez_2011.txt (L-lylh) EL SINTAGMA NOMINAL EN LA EXTRACCIÓN DE RELACIONES LÉXICO SEMÁNTICAS DE CONTEXTOS DEFINITORIOS EL CASO DE LA PREPOSICIÓN DE	Eadrianareyescareaga_2008.txt (L-lylh) REGLAS DE CORRESPONDENCIA ENTRE SONIDO Y GRAFÍA EN EL ESPAÑOL HABLADO EN MÉXICO EN EL SIGLO XVI PARA LA CREACIÓN DE UN TRANSCRIPTOR AUTOMÁTICO UNA APORTACIÓN AL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO CHEM Ejoseluisvieyra.txt (L-ingcomp) Adaptación optimización y expansión de Ecode un sistema extractor de contextos definitorios Eazuryaparicioaguilar_2011.txt (L-Enfermería) ESTADO DEL CONOCIMIENTO SOBRE SEXUALIDAD HUMANA SALUD SEXUAL Y SALUD REPRODUCTIVA UN ESTUDIO BIBLIOMÉTRICO DE LA PRODUCTIVIDAD CIENTÍFICA EN ESPAÑOL DEL 2001 2010 Erodrigoalarconmartinez_2009.txt (DR) Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios Eolgaacostalopez_2013.txt (D-ingcomp) EXTRACCIÓN AUTOMÁTICA DE RELACIONES LÉXICO SEMÁNTICAS A PARTIR DE TEXTOS ESPECIALIZADOS Eoctaviosanchezvelazquez_2009.txt (L-lylh) LA FUNCIONALIDAD AL INTERIOR DE CONTEXTOS DEFINITORIOS CON DEFINICIONES ANALÍTICAS EL PATRÓN SINTÁCTICO PARA INFINITIVO Ealejandromolinavillegas_2009.txt (M-ciencomp) AGRUPAMIENTO SEMÁNTICO DE CONTEXTOS DEFINITORIOS	Egabrielcastillohernandez_2012.txt (M-ciencomp) ALGORITMO REVISADO PARA LA EXTRACCIÓN AUTOMÁTICA DE AGRUPAMIENTOS SEMÁNTICOS Eantonioreyesperez_2006.txt (M-lylh) ESTRUCTURACIÓN SEMÁNTICO PRAGMÁTICA DEL LÉXICO EN DOMINIOS RESTRINGIDOS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

No. Experimento: 14		
0	1	2
Ecesarantonioaguilar_2008.txt (D-Ling) Análisis lingüístico de definiciones en contextos Definitorios Eirasemacruzdominguez_2011.txt (L-lylh) EL SINTAGMA NOMINAL EN LA EXTRACCIÓN DE RELACIONES LÉXICO SEMÁNTICAS DE CONTEXTOS DEFINITORIOS EL CASO DE LA PREPOSICIÓN DE Ecarlosfranciscomendez_2009.txt (M-Ling) IDENTIFICACIÓN AUTOMÁTICA DE CATEGORÍAS GRAMATICALES EN ESPAÑOL DEL SIGLO XVI Emarinavladimirovnafomicheva_2012.txt (M-Ling) ANÁLISIS LINGÜÍSTICO DE LA TRADUCCIÓN AUTOMÁTICA PARA SU EVALUACIÓN Eximenagutierrezvazquez_2010.txt (L-ingcomp) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO Ealexgarduñosandoval_2010.txt (L-ingcomp) DESARROLLO DE UN CONSTRUCTOR AUTOMÁTICO DE BASES DE DATOS RELACIONALES A PARTIR DE ESQUEMAS XML Eadrianamirandanava_2006.txt (L-ingcomp) GEOVARIANTES LÉXICAS DEL ESPAÑOL Ecarlosfranciscomendez_2013.txt (D-Ling) GENERACIÓN AUTOMÁTICA DE UNA GRAMÁTICA DE ESTADOS FINITOS PARA LA MORFOLOGÍA DEL ESPAÑOL Eitacruzsherling_2013.txt (L-ingcomp) CONTEXTOS DEFINITORIOS EN LA EXTRACCIÓN DE TAXONOMÍAS EL CASO DE PLN Emaríajimenezvasques_2012.txt (L-ING) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO	Eluiscabreradiego_2011.txt (L-ingcomp) TF IDF PARA LA OBTENCIÓN AUTOMÁTICA DE TÉRMINOS Y SU VALIDACIÓN MEDIANTE WIKIPEDIA Esistemasrecuperacioninformacion.txt (M-Ling) ESTRUCTURACIÓN SEMÁNTICO PRAGMÁTICA DEL LÉXICO EN DOMINIOS RESTRINGIDOS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN Ebrendacastrorolon.txt (L-lylh) DETECCIÓN DE SIMILITUD TEXTUAL MEDIANTE CRITERIOS DE DISCURSO Y SEMÁNTICA Epavelsorionomorales_2011.txt (L-ingcomp) CLASIFICACIO N DE OPINIONES MEDIANTE APRENDIZAJE DE MA QUINAS EL CASO DE RESEN AS SOBRE PELI CULAS Eariadnahernandezanguilo_2009.txt (L-lylh) ANÁLISIS LINGÜÍSTICO DE DEFINICIONES ANALÍTICAS PARA LA BÚSQUEDA DE REGLAS QUE PERMITAN SU DELIMITACIÓN AUTOMÁTICA Eoctaviosanchezvelazquez_2009.txt (L-lylh) LA FUNCIONALIDAD AL INTERIOR DE CONTEXTOS DEFINITORIOS CON DEFINICIONES ANALÍTICAS EL PATRÓN SINTÁCTICO PARA INFINITIVO Egabrielcastillohernandez_2012.txt (M-ciencomp) ALGORITMO REVISADO PARA LA EXTRACCIÓN AUTOMÁTICA DE AGRUPAMIENTOS SEMÁNTICOS	Epérezpéreznuri.txt (M-Ling) PROBLEMAS DE LOS DICCIONARIOS ESPECIALIZADOS deltutado DE LIBRE COMERCIO DE AMÉRICA DEL NORTE CASO práctico SOBRE LA terminología REFERENTE A LA inversión EXTRANJERA Eazuryaparicioaguilar_2011.txt (L-Enfermería) ESTADO DEL CONOCIMIENTO SOBRE SEXUALIDAD HUMANA SALUD SEXUAL Y SALUD REPRODUCTIVA UN ESTUDIO BIBLIOMÉTRICO DE LA PRODUCTIVIDAD CIENTÍFICA EN ESPAÑOL DEL 2001 2010 Ejesusvaldesramos_2013.txt (D-Bibliotec) BASES LINGÜÍSTICAS E INFORMÁTICAS PARA LA ELABORACIÓN DE TESAUROS

No. Experimento: 16		
0	1	2
Ejuliovargasmejia_2013.txt (L-Inf) OPTIMIZACIÓN DEL GENERADOR DE CONCORDANCIAS DEL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO CHEM MEDIANTE UNA COMPARACIÓN ENTRE MANEJADORES DE BASES DE DATOS RELACIONALES Y ADMINISTRACIÓN DE DESEMPEÑO DE BASES DE DATOS Eadrianamirandanava_2006.txt (L-ingcomp) GEOVARIANTES LÉXICAS DEL ESPAÑOL Ealejandrorosas_2011.txt (L-lylh) ANÁLISIS ESTILOMÉTRICO PARA LA DETECCIÓN DE PLAGIO Eirasemacruzdominguez_2011.txt (L-lylh) EL SINTAGMA NOMINAL EN LA EXTRACCIÓN DE RELACIONES LÉXICO SEMÁNTICAS DE CONTEXTOS DEFINITORIOS EL CASO DE LA PREPOSICIÓN DE Eantoniaestradatrejo_2009.txt (L-lylh) ANÁLISIS DE LAS RELACIONES CONCEPTUALES EN UNA TERMINOLOGÍA APLICACIÓN AL DICCIONARIO DE LINGÜÍSTICA Elauraelenahernandez.txt (M-cienccomp) CREACIÓN SEMI AUTOMÁTICA DE LA BASE DE DATOS Y MEJORA DEL MOTOR DE BÚSQUEDA DE UN DICCIONARIO ONOMASIOLÓGICO Evaleriabenitezrosete_2008.txt (L-lylh) ANÁFORAS EN LA EXPANSIÓN DE CONTEXTOS DEFINITORIOS UNA PROPUESTA DE ETIQUETADO Ejesusvaldesramos_2013.txt (D-Bibliotec) BASES LINGÜÍSTICAS E INFORMÁTICAS PARA LA ELABORACIÓN DE TESAUROS Evictorsanchez-marcovelazquez_2013.txt (L-ingcomp) DESARROLLO DE UN EXTRACTOR DE PALABRAS CLAVE Epérezpéreznuri.txt (M-Ling) PROBLEMAS DE LOS DICCIONARIOS ESPECIALIZADOS deltutado DE LIBRE COMERCIO DE AMÉRICA DEL NORTE CASO práctico SOBRE LA terminología	Esandrarichermonroy_2010.txt (L-lylh) ELABORACIÓN DE UN CORPUS ETIQUETADO DE DISCURSO INFANTIL ESCRITO Esistemasrecuperacioninformacion.txt (M-Ling) ESTRUCTURACIÓN SEMÁNTICO PRAGMÁTICA DEL LÉXICO EN DOMINIOS RESTRINGIDOS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN Eximenagutierrezvazquez_2010.txt (L-ingcomp) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO Emaríajimenezvasques_2012.txt (L-ING) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO	Ecarlosfranciscomendez_2009.txt (M-Ling) IDENTIFICACIÓN AUTOMÁTICA DE CATEGORÍAS GRAMATICALES EN ESPAÑOL DEL SIGLO XVI

REFERENTE A LA inversión EXTRANJERA Eolgaacostalopez_2013.txt (D-ingcomp) EXTRACCIÓN AUTOMÁTICA DE RELACIONES LÉXICO SEMÁNTICAS A PARTIR DE TEXTOS ESPECIALIZADOS Ejorgelazarohernandez_2010.txt (L-lylh) EXTRACCIÓN DE LA TERMINOLOGÍA BÁSICA DE LAS SEXUALIDADES EN MÉXICO A PARTIR DE UN CORPUS LINGÜÍSTICO Eedgarmoralespalafox_2011.txt (M-cienccomp) DESARROLLO DE UN MODELO COMPUTACIONAL PARA LA SOLUCIÓN DE BLOQUEOS MEDIANTE ANALOGÍAS EN EL SISTEMA MEXICA Emarinavladimirovnafomicheva_2012.txt (M-Ling) ANÁLISIS LINGÜÍSTICO DE LA TRADUCCIÓN AUTOMÁTICA PARA SU EVALUACIÓN		

No. Experimento: 18		
0	1	2
Esistemasrecuperacioninformacion.txt (M-Ling) ESTRUCTURACIÓN SEMÁNTICO PRAGMÁTICA DEL LÉXICO EN DOMINIOS RESTRINGIDOS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN	Esandrarichermonroy_2010.txt (L-lylh) ELABORACIÓN DE UN CORPUS ETIQUETADO DE DISCURSO INFANTIL ESCRITO Eirasemacruzdominguez_2011.txt (L-lylh) EL SINTAGMA NOMINAL EN LA EXTRACCIÓN DE RELACIONES LÉXICO SEMÁNTICAS DE CONTEXTOS DEFINITORIOS EL CASO DE LA PREPOSICIÓN DE Ecarlosfranciscomendez_2009.txt (M-Ling) IDENTIFICACIÓN AUTOMÁTICA DE CATEGORÍAS GRAMATICALES EN ESPAÑOL DEL SIGLO XVI	Eximenagutierrezvazquez_2010.txt (L-ingcomp) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO Emaríajimenezvasques_2012.txt (L-ING) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO Ejuliovargasmejia_2013.txt (L-Inf) OPTIMIZACIÓN DEL GENERADOR DE CONCORDANCIAS DEL CORPUS HISTÓRICO DEL ESPAÑOL EN MÉXICO CHEM MEDIANTE UNA COMPARACIÓN ENTRE MANEJADORES DE BASES DE DATOS RELACIONALES Y ADMINISTRACIÓN DE DESEMPEÑO DE BASES DE DATOS Eadrianamirandanava_2006.txt (L-ingcomp) GEOVARIANTES LÉXICAS DEL ESPAÑOL Ealejandrorosas_2011.txt (L-lylh) ANÁLISIS ESTILOMÉTRICO PARA LA DETECCIÓN DE PLAGIO Eantoniaestradatrejo_2009.txt (L-lylh) ANÁLISIS DE LAS RELACIONES CONCEPTUALES EN UNA TERMINOLOGÍA APLICACIÓN AL DICCIONARIO DE LINGÜÍSTICA Elauraelenahernandez.txt (M-cienccomp) CREACIÓN SEMI AUTOMÁTICA DE LA BASE DE DATOS Y MEJORA DEL MOTOR DE BÚSQUEDA DE UN DICCIONARIO ONOMASIOLÓGICO Evaleriabenitezrosete_2008.txt (L-lylh) ANÁFORAS EN LA EXPANSIÓN DE CONTEXTOS DEFINITORIOS UNA PROPUESTA DE ETIQUETADO Ejesusvaldesramos_2013.txt (D-Bibliotec) BASES LINGÜÍSTICAS E INFORMÁTICAS PARA LA ELABORACIÓN DE TESAUROS Evictorsanchez-marcovelazquez_2013.txt (L-ingcomp) DESARROLLO DE UN EXTRACTOR DE PALABRAS CLAVE Epérezpéreznuri.txt (M-Ling) PROBLEMAS DE LOS DICCIONARIOS ESPECIALIZADOS DEL TRATADO DE

		LIBRE COMERCIO DE AMÉRICA DEL NORTE CASO práctico SOBRE LA terminología REFERENTE A LA inversión EXTRANJERA Eolgaacostalopez_2013.txt (D-ingcomp) EXTRACCIÓN AUTOMÁTICA DE RELACIONES LÉXICO SEMÁNTICAS A PARTIR DE TEXTOS ESPECIALIZADOS Ejorgelazarohernandez_2010.txt (L-lylh) EXTRACCIÓN DE LA TERMINOLOGÍA BÁSICA DE LAS SEXUALIDADES EN MÉXICO A PARTIR DE UN CORPUS LINGÜÍSTICO Eedgarmoralespalafox_2011.txt (M-cienccomp) DESARROLLO DE UN MODELO COMPUTACIONAL PARA LA SOLUCIÓN DE BLOQUEOS MEDIANTE ANALOGÍAS EN EL SISTEMA MEXICA Emarinavladimirovnafomicheva_2012.txt (M-Ling) ANÁLISIS LINGÜÍSTICO DE LA TRADUCCIÓN AUTOMÁTICA PARA SU EVALUACIÓN

No. Experimento: 21		
0	1	2
Ejuliovargasmejia_2013.txt (L-Inf) Optimización del generador de concordancias del corpus histórico del español en méxico chem mediante una comparación entre manejadores de bases de datos relacionales y administración de desempeño de bases de datos Eadrianamirandanava_2006.txt (L-ingcomp) GEOVARIANTES LÉXICAS DEL ESPAÑOL Ealejandrorosas_2011.txt (L-lylh) ANÁLISIS ESTILOMÉTRICO PARA LA DETECCIÓN DE PLAGIO Eirasemacruzdominguez_2011.txt (L-lylh) EL SINTAGMA NOMINAL EN LA EXTRACCIÓN DE RELACIONES LÉXICO SEMÁNTICAS DE CONTEXTOS DEFINITORIOS EL CASO DE LA PREPOSICIÓN DE Eantoniaestradatrejo_2009.txt (L-lylh) ANÁLISIS DE LAS RELACIONES CONCEPTUALES EN UNA TERMINOLOGÍA APLICACIÓN AL DICCIONARIO DE LINGÜÍSTICA Elauraelenahernandez.txt (M-cienccomp) CREACIÓN SEMI AUTOMÁTICA DE LA BASE DE DATOS Y MEJORA DEL MOTOR DE BÚSQUEDA DE UN DICCIONARIO ONOMASIOLÓGICO Evaleriabenitezrosete_2008.txt (L-lylh) ANÁFORAS EN LA EXPANSIÓN DE CONTEXTOS DEFINITORIOS UNA PROPUESTA DE ETIQUETADO Ejesusvaldesramos_2013.txt (D-Bibliotec) BASES LINGÜÍSTICAS E INFORMÁTICAS PARA LA ELABORACIÓN DE TESAUROS Evictorsanchez-marcovelazquez_2013.txt (L-ingcomp) DESARROLLO DE UN EXTRACTOR DE PALABRAS CLAVE Epérezpéreznuri.txt (M-Ling) PROBLEMAS DE LOS DICCIONARIOS ESPECIALIZADOS deltutado DE LIBRE COMERCIO DE AMÉRICA DEL NORTE CASO práctico SOBRE LA terminología	Esandrarichermonroy_2010.txt (L-lylh) ELABORACIÓN DE UN CORPUS ETIQUETADO DE DISCURSO INFANTIL ESCRITO Esistemasrecuperacioninformacion.txt (M-Ling) ESTRUCTURACIÓN SEMÁNTICO PRAGMÁTICA DEL LÉXICO EN DOMINIOS RESTRINGIDOS PARA SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN Eximenagutierrezvazquez_2010.txt (L-ingcomp) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO Emaríajimenezvasques_2012.txt (L-ING) SISTEMA DE RESUMEN EXTRACTIVO AUTOMÁTICO	Ecarlosfranciscomendez_2009.txt (M-Ling) IDENTIFICACIÓN AUTOMÁTICA DE CATEGORÍAS GRAMATICALES EN ESPAÑOL DEL SIGLO XVI

REFERENTE A LA inversión EXTRANJERA Eolgaacostalopez_2013.txt (D-ingcomp) EXTRACCIÓN AUTOMÁTICA DE RELACIONES LÉXICO SEMÁNTICAS A PARTIR DE TEXTOS ESPECIALIZADOS Ejorgelazarohernandez_2010.txt (L-lylh) EXTRACCIÓN DE LA TERMINOLOGÍA BÁSICA DE LAS SEXUALIDADES EN MÉXICO A PARTIR DE UN CORPUS LINGÜÍSTICO Eedgarmoralespalafox_2011.txt (M-cienccomp) DESARROLLO DE UN MODELO COMPUTACIONAL PARA LA SOLUCIÓN DE BLOQUEOS MEDIANTE ANALOGÍAS EN EL SISTEMA MEXICA Emarinavladimirovnafomicheva_2012.txt (M-Ling) ANÁLISIS LINGÜÍSTICO DE LA TRADUCCIÓN AUTOMÁTICA PARA SU EVALUACIÓN		

Anexo 8

Gazeteers elaborados para el reconocimiento de elementos en los textos extraídos.

Puesto de trabajo:

- director
- directora
- secretaria académica
- secretario académico
- secretaria administrativa
- secretario administrativo
- secretaria técnica
- secretario técnico
- secretaria de planeación y desarrollo académico
- secretario de planeación y desarrollo académico
- subdirector de hidráulica y ambiental
- subdirector de estructuras y geotecnia
- subdirector de electromecánica
- subdirectora de hidráulica y ambiental
- subdirectora de estructuras y geotecnia
- subdirectora de electromecánica
- coordinador de ingeniería ambiental
- coordinador de hidráulica
- coordinador de ingeniería de procesos industriales y ambientales
- coordinador de mecánica aplicada
- coordinador de geotecnia
- coordinador de estructuras y materiales
- coordinador de ingeniería sismológica
- coordinador de instrumentación sísmica
- coordinador de mecánica y energía
- coordinador de eléctrica y computación
- coordinador de ingeniería en sistemas
- coordinador de sistemas de cómputo
- coordinador de electrónica
- coordinadora de ingeniería ambiental
- coordinadora de hidráulica
- coordinadora de ingeniería de procesos industriales y ambientales
- coordinadora de mecánica aplicada
- coordinadora de geotecnia
- coordinadora de estructuras y materiales
- coordinadora de ingeniería sismológica
- coordinadora de instrumentación sísmica
- coordinadora de mecánica y energía
- coordinadora de eléctrica y computación
- coordinadora de ingeniería en sistemas
- coordinadora de sistemas de cómputo

- coordinadora de electrónica
- jefe de juríquilla
- jefe de sisal
- jefa de juríquilla
- jefa de sisal
- subdirección
- coordinación
- unidad académica juríquilla
- unidad académica sisal
- departamento de contabilidad
- departamento de bienes y suministros
- departamento de personal
- departamento de presupuesto
- departamento de servicios generales
- unidad de gestión de convenios y contratos
- unidad de servicios de información
- unidad de docencia y formación de recursos humanos
- unidad de patentes y transferencia tecnológica
- unidad de promoción y comunicación
- unidad de informática y control estadístico
- unidad de apoyo a cuerpos colegiados
- unidad de gestión de financiamiento
- unidad de instrumentación sísmica
- área de contabilidad
- área de ingresos extraordinarios
- comité interno
- comité dictaminador
- comité dictaminadora
- comité página web
- administración y planeación
- publicaciones
- subcomité de adquisiciones
- grupo de planeación
- dirección
- investigador titular
- investigador asociado
- investigador emérito
- técnico académico titular
- técnico académico asociado
- sistemas de cómputo en bases de datos
- sistemas de cómputo en capacitación
- sistemas de cómputo en pc's
- sistemas de cómputo en redes
- sistemas de cómputo en servicios linux
- sistemas de cómputo en servicios windows
- sistemas de cómputo en sitios web

- departamento

Periodicidad:

- diaria
- semanal
- mensual
- bimestral
- trimestral
- cuatrimestral
- semestral
- anual
- bianual
- a demanda

Grado académico:

- licenciatura
- maestría
- doctorado

Nombre de sección de tesis:

- metodología
- capítulo
- hipótesis
- preguntas de investigación
- tutor
- autor
- resultados
- experimentos

Nombre de archivo de expediente:

- programa de trabajo
- informe de actividades
- comprobante de inscripción
- historial académico

Referencias

Aramburo, M. B. (1999). El conocimiento como fuente de ventaja competitiva. *La gestión de la diversidad: XIII Congreso Nacional, IX Congreso Hispano-Francés*, 485-490.

Asian Productivity Organization. (2009). *Knowledge Management: Facilitarors' Guide.* (P. Nair, & K. Prakash, Edits.) Tokio.

Asian Productivity Organization. (2010). *Knowledge Management Tools and Techniques Manual.* (R. Young, Ed.) Tokyo, Japón: Asian Productivity Organization.

Badia Cardús, T. (2003). Técnicas de procesamiento del lenguaje. En M. A. Martí Antonín, *Tecnologías del lenguaje* (págs. 193-245). Barcelona: UOC.

Beesley, L., & Cooper, C. (2008). Defining knowledge management (KM) activities: Towards consensus. *Journal of Knowledge Management*, *12*, 48-62.

Benavides Velazco, C. A., & Quintana García, C. (2003). *Gestión del conocimiento y calidad total.* Madrid: Díaz de Santos.

Benbya, H., Passiante, G., & Aissa, N. (2004). Corporate portal: a tool for knowledge management synchronization. *International Journal of Information Management*, 201-220.

Biggam, J. (2001). Defining Knowledge: an Epistemological Foundation for Knowledge Management. *Proceedings of the 34th Hawaii International Conference on System Sciences.* Hawaii: ieee.

Borst, P., & Akkermans, H. (1997). An ontology approach to product disassembly. *10th European Workshop, EKAW '97*, (págs. 33-48). Catalonia, España.

Canals, A. (2003). *Gestión del conocimiento.* Barcelona: Gestión 2000.

Centro de sistemas de conocimiento, Tecnológico de Monterrey. (2001). *Administración del conocimiento en México: Entendimiento, intención, práctica, resultados y visión a futuro.* Monterrey, México.

Chandra, A., & Khanijo, M. (2009). Participation of the international management institute in the knowledge economy project. *Second international conference on technology and innovation for knowledge management* (págs. 101-106). Japan: Asian Productivity Organization.

Chinchor, N., Lewis, D., & Hirschman, L. (1993). Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3). *Computational linguistics*, *19*(3), 409-449.

CIDEC. Centro de Investigación y Documentación sobre problemas de la Economía, el Empleo y las Cualificaciones Profesionales. (2004). *Gestión del conocimiento y capital intelectual* (Vol. 31). Donostia - San Sebastián, España: Gobierno Vasco, Eusko Jaurlaritza.

Ciravegna, F. (2001). Challenges in Information Extraction from Text for Knowledge Management. *IEEE Intelligent Systems and Their Applications.* IEEE.

Diao, L., Zuo, M., & Liu, Q. (2009). The Artificial Intelligence in Personal Knowledge Management. *Second International Symposium on Knowledge Acquisition and Modeling* (págs. 327- 329). IEEE.

Dias, C. (2001). Corporate Portals: a literature review of a new concept in Information Management. *International Journal of Information Management*, 270-287.

Drucker, P. F., & Centre Canadien de gestion. (1995). The age of social transformation.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37-51.

Feldman, R., & Dagan, I. (1995). Knowledge iscovery in textual databases (KDT). *KDD-95*, (págs. 112-117).

Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., & Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *2nd International Conference on Practical Aspects of Knowledge Management. 13*, págs. 1-10. Basel, Switzerland: U. Reimer.

Feng, Y. (2010). Towards Knowledge Discovery in Semantic Era. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. 5*, págs. 2071-2075. IEEE.

Ganesh D., B. (2001). Knowledge management in organizations: examining the interaction between technologies, techniques, and people. *Journal Of Knowledge Management*, 68-75.

Garud, R., & Kumaraswamy, A. (2005). Vicious and Virtuous Circles in the Management of Knowledge: The Case of Infosys Technologies. *Management Information Systems Research Center, University of MinnesotaStable*, 9-33.

Gianneto, K., & Wheeler, A. (2002). *Herramientas para la administración del capital intelectual.* Panorama.

Gibert, K. (2004). Técnicas híbridas de inteligencia artificial para el descubrimiento de conocimiento y la minería de datos. *Tendencias de la minería de datos en España*, 119-130.

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques.* Springer.

Grau, A. (Octubre de 2003). *Fundación Iberoamericana del Conocimiento.* Obtenido de http://www.gestiondelconocimiento.com

Grishman, R. (2010). Information Extraction. En A. Clark, C. Fox, & S. Lappin, *The handbook ok Computational Linguistics and Natural Language Processing* (págs. 517- 530). United Kingdom: Wiley-Blackwell.

Gupta, V., & Lehal, G. S. (Agosto de 2009). A survey of Text Mining Techniques and Applications. *Journal of emerging technologies in web intelligente, 1*(1), 60-76.

Hartigan, J. A., & Manchek, W. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society*, 100-108.

Hearts, M. A. (1999). Untangling text data mining. *37th annual meeting of the Association for Computational Linguistics* (págs. 3-10). Association for Computational Linguistics.

Heisig, P. (2009). Harmonisation of knowledge management – comparing 160 KM frameworks around the globe. *JOURNAL OF KNOWLEDGE MANAGEMENT, 13*(4), 4-31.

Hotho, A., Nürnberger, A., & Paab, G. (2005). A brief survey of text mining. *LDV Forum*, 19-62.

Hume, D. (1957). Enquiries Concerning the Human Understanding: And Concerning the Principles of Morals. (S. Bigge, Ed.)

Jennex, M. E., Olfman, L., & Addo, T. B. (2002). The need for an organizational knowledge management strategy. *Proceedings of the 36th Hawaii International Conference on System Sciences.* Hawaii: IEEE.

Kodratoff, Y. (1999). Knowledge discovery in texts: a definition, and applications. *Foundations of Intelligent Systems*, 16-29.

Kotter, J. P. (1996). *Leading change.* Harvard Business Press.

Kremer, S., Smolnik, S., & Kolbe, L. (2004). Towards Knowledge Discovery through Context Explication. *37th Hawaii International Conference on Systems Sciences.* Hawaii.

Krüger, K. (2006). El concepto de sociedad del conocimiento. *Revista bibliográfica de geografía y ciencias sociales*(683).

Linkage. (2000). Inc.'s best practices in knowledge management and organizational learning handbook: Case studies, instruments, models, research. *Linkage*.

Linlin, J., & Hui, Z. (2008). An empirical study on the knowledge management elements of research team in university. *IEEE*.

Liu, B. (2007). *Web data mining.* Springer.

Liu, Y., Liu, X.-h., & Yang, A.-g. (2008). The Application and Research of Ontology in Knowledge Management Field. *2008 IFIP International Conference on Network and Parallel Computing*, (págs. 561-564).

Llisterri, J. (2003). Lingüística y tecnologías del lenguaje. *Lynx. Panorámica de Estudios Lingüísticos*(2), 9-71.

Loebbecke, C., & Crowston, K. (2012). Knowledge Portals: Components, Functionalities, and Deployment Challenges.

Maedche, A. (2003). *Ontology learning for the semantic web.* Massachusetts: Kluwer Academic Publishers.

Maedche, A., Motik, B., Stojanovic, L., Studer, R., & Volz, R. (2003). Ontologies for Enterprise Knowledge Management. *IEEE Inteligent systems*, 26-33.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge university press.

Mansingh, G., Osei-Bryson, K.-M., & Reichgelt, H. (2009). Building ontology-based knowledge maps to asist knowledge process outsourcing decisions. *Knowledge Management Research and Practice*, 37-51.

Martí Antonín, M. A., Alonso Martín, J. A., Badia Cardús, T., Campás Montaner, J., Gómez Guinovart, X., Gonzalo Arroyo, J., . . . Verdejo Maíllo, M. F. (2003). *Tecnologías del lemguaje.* Barcelona: UOC.

Martínez Sánchez, A., & Corrales Estrada, M. (2011). *Administración del conocimiento y desarrollo basado en conocimiento.* México DF: Cengage Learning.

Maybury, M. (2001). Human Language Technologies for Knowledge Management: Challenges and Opportunities. *HLTKM '01 Proceedings of the workshop on Human Language Technology and Knowledge Management.* Association for Computational Linguistics.

Moldovan, D. (2001). Question-Answering systems in knowledge management. IEEE.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 47-60.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. En *The strategic management of intellectual capital and organizational knowledge* (págs. 437-462).

Ordoñez De Pablos, P. (2001). La gestión del conocimiento como base para el logro de una ventaja competitiva sostenible: La organización occidental versus japonesa. *Investigaciones Europeas de Dirección y economía de la Empresa, 7*(3), 91-108.

Paz López, M. E., López Molina, E., & Solórzano Rodas, J. (2011). *¿El conocimiento, principal fuente de la globalizacion?* (primera ed.). México: Universidad Autónoma Benito Juárez de Oaxaca, Porrua.

Polanyi, M. (1967). The tacit dimension. (Routledge, & P. Kegan, Edits.)

Porter, M. E. (1995). *Ventaja competitiva: creación y sostenimiento de un desempeño superior.* Rei Argentina.

Redner, R. A., & Walker, H. F. (Abril de 1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *Society for Industrial and Applied Mathematics, 26*(2), 195-239.

Rivera, O. (2000). La Gestión del Conocimiento en el mundo Académico: ¿Cómo es la universidad de la era del conocimiento? *AECA.*

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management, 24*(5), 513-523.

Sánchez Medina, A. J., Melián González, A., & Hormiga Pérez, E. (2007). El concepto de capital intelectual y sus dimensiones. *Investigaciones Europeas de Dirección y Economía de la Empresa*, *13*(2), 97-111.

Secretaría General. Universidad Nacional Autónoma de México. (10 de Enero de 2012). *Dirección General de Cómputo y de Tecnologías de Información y Comunicación.* Recuperado el 24 de Noviembre de 2014, de http://www.tic.unam.mx/acerca.html

Sierra Martínez, G. (2009). Visión interdisciplinaria en la prospectiva de las tecnologías del lenguaje en México. En J. C. Villa Soto, *El dominio de la lingüística* (págs. 135-160). México D.F.: Universidad Nacional Autónoma de México.

Sierra, G. E. (2006). Perspectivas de un Centro Nacional de Conocimiento, Información y Tecnologías del Lenguaje. *Tecnológica 2001*, 3.

Staab, S. (2001). Human language technologies for knowledge management. *IEEE Intelligent systems*, 84-94.

Stewart, T., & Ruckdeschel, C. (1998). Intellectual capital: The new wealth of organizations. *Performance Improvement*, 56-59.

Supyuenyong, V., & Islam, N. (2006). Knowledge Management Architecture: Building Blocks and Their Relationships. *PICMET*, (págs. 1210-1219). Istanbul, Turrkey.

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. *PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, (págs. 65-70).

Tedmori, S., & Jackson, T. W. (2012). The design and evaluation of EKE, a semi-automated email knowledge extraction tool. *Knowledge Management Research and Practice*, 79-88.

Tyndale, P. (2002). A taxonomy of knowledge management software tools: origins and applications. *Evaluation and Program Planning*(25), 183-190.

Valerio, G. (2002). Herramientas tecnológicas para la administración del conocimiento. *Transferencia, año 15*(57), 19-21.

Vorakulpipat, C., & Rezgui, Y. (2008). Value creation: the future of knowledge management. *The Knowledge Engineering Review*, 283-294.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2005). *Text Mining. Predictive Methods for Analyzing Unstructured Information.* New York: Springer.

Yang, W., Chang-xiong, S., & Xue-mei, D. (2006). Constructing models of knowledge management in research teams. *ICMSE '06 International conference on* (págs. 1360-1365). Lille: Management science ang engineering.

Yang, W., Chang-xiong, S., Lei, Z., Li-yan, M., & Ying, J. (2008). Constructing the application models of knowledge management and innovation based on communication means in research team. *IEEE*.

Yang, W., Jing-Jun, Z., & Chang-xiong, S. (2007). The Fusion Model of Knowledge Management and Communication Management in Research Organization. *ieee*.