***ANEXO 1***

# 1 Revisión Sistemática

En este ANEXO se hablara de la metodología de investigación utilizada y las modificaciones a la misma de manera que se adapte al desarrollo de este proyecto de fin de carrera. Se iniciara ahondando sobre los principios básicos de la revisión sistemática para luego proponer una plantilla que podría servir de input en futuras investigaciones relacionadas.

## 1.1 Descripción de la investigación

Como ya se explicó en el marco conceptual la revisión sistemática nos permite realizar una investigación ordenada y esquemática con la finalidad de poder copiar y mejorar este proceso en diferentes proyectos.

Esta metodología se adaptó para este proyecto de fin de carrera utilizando otros métodos que permitieron establecer la dirección del mismo. Entre las metodologías utilizadas en esta revisión se encuentra el diseño de un árbol de problemas acompañado del método Blaxter con preguntas de investigación.

### *1.1.1 Preguntas de Investigación*

¿Qué? ¿Cómo? ¿Cuándo? ¿Por qué?, a manera general estas 4 interrogantes siempre ayudan a encontrar respuestas simples a problemas complejos ya que nos definen una gran cantidad de aspectos, es por ello que para realizar esta investigación se identificó una serie de preguntas clave a manera que sean resueltas en desarrollo de la misma y nos sirvan como guía a la horas de realizar la revisión de la literatura, cabe resaltar que para la realización de las preguntas de investigación se desarrollaron en conjunto con el asesor debido a las complicaciones del tema y a la expertis del mismo.

| Relacion | Pregunta | Concepto Base |
|---|---|---|
| Marco Conceptual | ¿Qué es una Ontologia? | Ontolgias |
| | ¿Cómo se utiliza una ontología para hacer busquedas semanticas basadas en conocimiento? | Ontolgias |
| | ¿Cómo la recuperación de información soporta la recuperación de documentos? | Recuperacion de Informacion |
| | ¿Cuál es la ventaja de la expansión de consulta fente a otras técnicas? | Expansion de Consulta |
| | ¿Cuáles son las ventajas de las ontologias frente a las bases de datos convencionales? | Ontolgias |
| | ¿Cuáles son las ventajas de la recuperacion de información frente a la recuperacion de informacion via sql? | Recuperacion de Informacion |
| | ¿De que manera se puede enriquecer un documento en formato PDF,WORD,etc para facilitar la busqueda de su contenido? | Etiquetacion Semantica |
| | ¿Qué estructurasse utilizan para facilitar la recuperacion de información? | Etiquetacion Semantica |
| Estado del arte | ¿Qué trabajos han utilizado expansion de consulta como base? | Expansion de Consulta |
| | ¿Qué trabajos han utilizado Ontologias como base? | Ontolgias |
| | ¿Qué trabajos se han realizado en el area de Information Retrieval? | Recuperacion de Informacion |
| | ¿Cuales fueron las arquitecturas propuestas? | Ontolgias |
| | ¿Como fue realizada la expansion de consulta? (Heuristico) | Expansion de Consulta |
| | ¿En qué dominio se realizó la aplicación? | Ontolgias |
| | Jenna como framework para el manejo de ontologías ¿Variantes? | Recuperacion de Informacion |

*Tabla i: Preguntas de Investigación*

### *1.1.2 Árbol de problemas*

El método árbol de problemas fue utilizado debido a que nos brinda un amplio entendimiento de la problemática a resolver. En el que se expresan una relación entre causas y efectos que nos llevan a percibir de una manera más sencilla la problemática.

La secuencia de pasos con la que se trabajo fue la identificación de un problema en general que luego fue derivando a sus posibles causas y estas en posibles problemas que se deberían tratar para llegar a la resolver el problema principal.

La cantidad de causas y efectos para cada problema difiere de la complejidad del mismo, sin embargo se trató de agrupar en un número aproximado a 4 causas que nos proporcionen un paso inicial para la formulación de nuestros principales objetivos.

Esto nos lleva fácilmente a la elección de nuestros objetivos generales y específicos que desprenden de cómo es que hemos desarrollado las causas del árbol de problemas y así convertirlo en un árbol de objetivos. A continuación se muestra el desarrollo del árbol de problemas para el entendimiento global de la problemática a tratar.

En él se identifican las siguientes partes:

- **Problema Central**: en el problema central se ubica la idea base de la problemática, la idea que engloba a las demás ideas a manera que sirva como inicio para desglosar e ir encontrando las diferentes causas y efectos

del mismo. Para este caso particular se escogió, dado el problema principal de la tesis, una idea que se relaciona con la dificultad de recuperar conocimiento de interés en documentos no estructurados.

- **Causas**: dado el problema central se comenzó a especular las posibles causas del problema central desde las más generales a específicas que nos sirvan como inicio para comenzar con la revisión de la literatura. En el caso de nuestra problemática se escogió lo siguiente: Contexto amplio, documentos desordenados/ falta de estándares, bases de datos convencionales no solucionan completamente el problema y la dificultad de elegir diferentes métodos de ordenamiento y recuperación.
- **Efectos**: son todos aquellos sucesos que se derivan del problema. En forma similar, se tiene que identificar los efectos *directos*, *indirectos* y *últimos* según su relación con el problema. Para efectos de este problema se delimito los relacionados a los usuarios y sus incomodidades a la hora de recuperar información.

A continuación se muestra el árbol de problemas desarrollado para este proyecto.

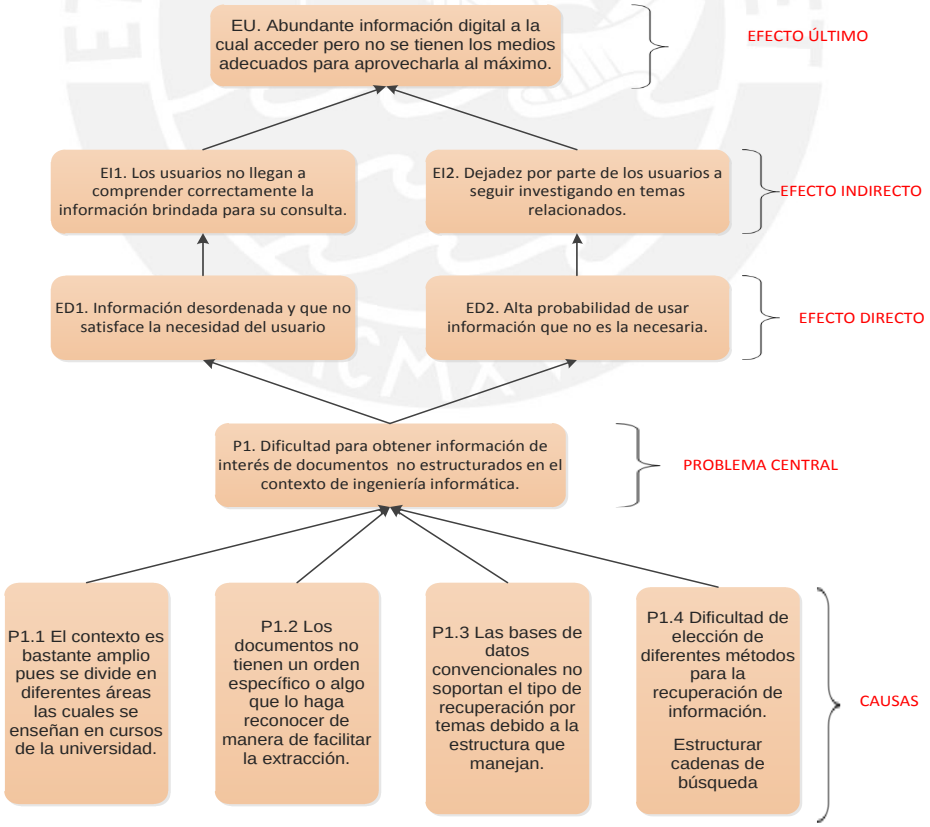


*Figura i: Árbol de problemas (Propia)*

### *1.1.3 Método Blaxter*

Loraine Blaxter en (Blaxter, Hughes, & Tight, 2006) nos menciona que actualmente las investigaciones cualitativas se caracterizan debido a que los hechos se deben interpretar en un contexto adecuado para que luego otro investigador puedo sumergirse, por así decirlo, en el ámbito seleccionado. Es por ello que veo la necesidad de contextualizar todas las ideas representadas en el árbol de problemas antes mencionado.

Cumpliendo con los principios fundamentales de una investigación cualitativa (Blaxter, Hughes, & Tight, 2006), se ha mapeado los resultados obtenidos con el árbol de problemas y así caracterizar correctamente el problema explicando adecuadamente las situaciones encontradas, además brindar contexto a las ideas a partir de la revisión de la literatura contestando las preguntas de la investigación.

| | Problema | Caracterización | Contexto | Concepto |
|---|---|---|---|---|
| Efecto Ultimo<br>Efecto Indirecto<br>Efecto Directo<br>Problema Central<br>Causa | Bibliografía | Explicación de la situación encontrada | Marco conceptual | Respuestas a las preguntas de investigación |

*Tabla ii: Matching Metodo Blaxter*

### *1.1.4 Revisión Sistemática*

Una revisión sistemática de la literatura permite identificar, evaluar, interpretar y sintetizar todas las investigaciones existentes y relevantes en un tema de interés particular. En este caso las investigaciones en el dominio de la Recuperación de Información usando ontologías son tomadas como input.

En (Kitchenham, 2004) Barbara Kitchenham presenta un método para la realización de revisiones sistemáticas en el contexto de Ingeniería de Software, la cual fue pionera en su clase ya que antes no se disponía de una guía o método eficiente para realizar estudios exhaustivos en ese dominio. El método propone 3 fases fundamentales *Planificación*, *Revisión* y *Publicación de Resultados.*

A continuación detallaré como se han realizado estas fases para el proyecto en desarrollo.

#### 1.1.4.1 Planificación

AL inicio del proyecto y en conjunto con el asesor se delimito el área a tratar de manera que permita tener una visión clara del tema de interés con la fijación de que este proceso sea iterativo en el transcurso del desarrollo de la tesis. En esta etapa se definió un *Protocolo de Revisión* donde se especificó que los temas a tratar eran los siguientes en orden de importancia en las búsquedas:

- Information Retrieval
- Ontologies
- Semantic Web
- Semantic Annotation
- Query Expansion

Además se establecieron ciertas bases de datos bibliográficas, que por su estilo y diseño son recomendados y confiables para la búsqueda de conocimiento entre las cuales tenemos:

- Scopus
- IEEE Xplore
- ACM
- ScienceDirect
- Google Scholar

Para la búsqueda de documentos con relación a los temas de interés se vio la necesidad de establecer ciertas k*eywords* o términos de búsqueda en la sintaxis de cada base de datos, entre las cuales tenemos los siguientes ejemplos:

| Buscador | Términos de Busqueda | N° de Articulos en 1er filtro |
|---|---|---|
| **SCOPUS** | TITLE("information retrieval" AND “ontology”) | 12 |
| **ACM** | Title: “information retrieval" and Title: “ontology” | 9 |
| **IEEE** | ( | 15 |

| | | |
|---|---|---|
| | ("Document Title": "information retrieval") AND ("Document Title": "ontology") ) | |
| **Science Direct** | TITLE(semantic web) and TITLE(ontology) TITLE(information retrieval) and TITLE(ontology) | 11 |
| **Google Scholar** | information retrieval + ontology Since 2000 semantic web + ontology Since 2000 | 8 |
| **USP** | Information retrieval + ontology | 15 |

*Tabla 1 Keywords para las busquedas*

Para centrarnos más en la investigación y tener una guía de búsqueda, se propuso una interrogante que abarca los temas a tratar según el enfoque PILOC (Petticrew & Roberts, 2005), la pregunta noteadora para este proyecto fue:

**¿Cómo se han utilizado las ontologías en el proceso de recuperación de conocimiento?**

En conjunto con las preguntas de investigación presentadas anteriormente se podía ya comenzar a realizar la revisión, pero además se planteó una estructura en la cual plasmar toda la información recopilada de manera que sea de fácil acceso para luego poder realizar las citaciones y demás usos, la estructura plantea un formulario de extracción de datos (Tranfield, Denyer, & Smart, 2003), es decir un conjunto de propiedades que ayudaran a identificar y realizar estadísticas al finalizar la revisión. Estas propiedades son:

- Nombre Articulo
- Dominio de la aplicación
- Propósito de la ontología
- Métodos, procedimiento y herramientas usadas
- Idioma
- Heurísticas
- Modelos planteados

- Tipo de información recuperada
- Año de la publicación
- Resultados

En el anexo “X” se aprecia un modelo de plantillas utilizadas que corresponden a la experiencia realizada y pueda ser reutilizada en trabajos futuros.

#### 1.1.4.2 Revisión

Luego de recopilar todos los documentos se utilizó la técnica de revisión veloz que consiste en leer solo los *abstracts* de los artículos de manera de enterarse, a grandes rasgos, del contenido del mismo y así realizar un primer filtro. Para luego clasificar los artículos según criterios de inclusión los cuales menciono a continuación:

- ¿El artículo se enfoca en recuperar conocimiento con el uso de ontologías? – Tier 1
- ¿El artículo presenta la utilización de ontologías para la representación de un dominio específico? – Tier 2
- ¿El artículo se enfoca en recuperar conocimiento con el uso de alguna otra tecnología? – Tier 3
- ¿El articulo habla sobre la importancia de obtener conocimiento pero no especifica los métodos a utilizar? – Tier 4

Para la selección de los artículos considerados primarios se definió como criterio de exclusión a todos los artículos de Tier 3 y 4.

Luego de aplicar dicho procedimiento se obtuvieron un total de 26 estudios filtrando un 50 % de la cantidad inicial de artículos recuperados. Todos estos se mapearon en el formulario de extracción de datos planteado en el anexo “X”.

#### 1.1.4.3 Publicación

A partir del formulario de extracción se pudo realizar estadísticas que abarcan diferentes aspectos.

- **Tendencias de las publicaciones:** desde la aparición de la WWW en 1998 se realizan diferentes estudios al respecto, recién a partir del 2000 Tim Bernners-Lee acuña el término de web semántica para denominar a un tipo de web en la que recuperar conocimiento sea más fácil, en este proyecto se tiene

conocimiento que la información que se tiene sobre web semántica es relativamente nueva y la información sobre las aplicaciones y uso de ontologías datan de los 90's es por ello que se puede observar el incremento de publicaciones a medida que pasan los años, incluyendo temas como etiquetación semántica y modelos de recuperación de conocimientos.

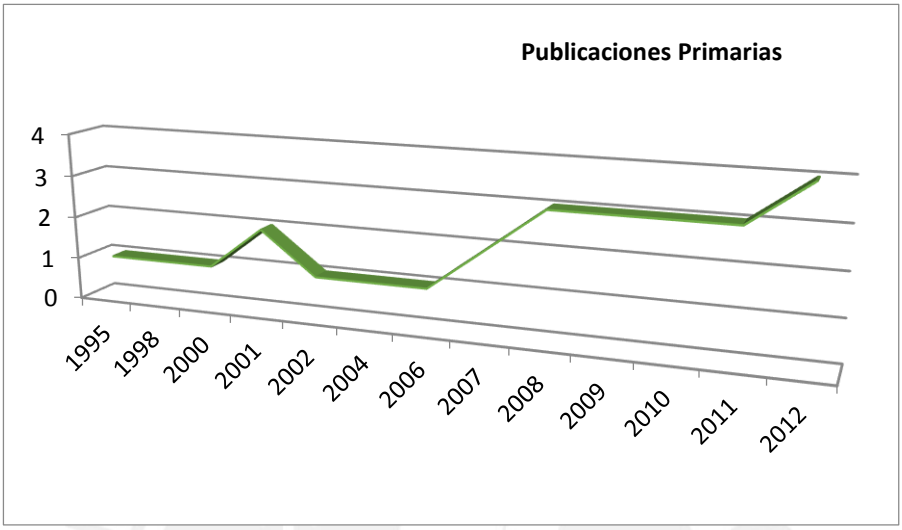


*Figura ii: Publicaciones primarias a lo largo de los años*

- **Bases de datos bibliográficas:** así mismo se observó que la mayoría de artículos utilizados en esta investigación se obtuvieron de los buscadores de la IEEE y de la USP (Universidad de San Paulo), esto debido a los temas que se buscaron y a que hablando geográficamente Brasil es uno de los países donde la investigación acerca de web semántica y knowledge managment está en su apogeo.

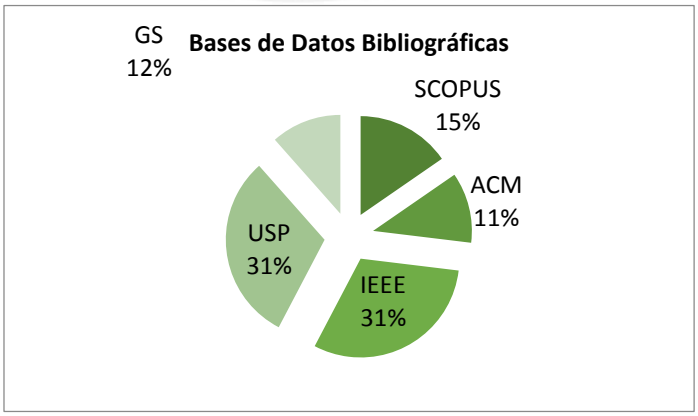


*Figura iii: Bases de datos bibliográficas*

- **Dominios utilizados:** en la actualidad es común encontrar diferentes modelos de conocimiento en diferentes dominios, pero el más común es encontrar en el ámbito de la medicina, lo cual tiene sentido ya que mapear el conocimiento en esa área para reemplazar la existencia de un experto es de suma importancia.

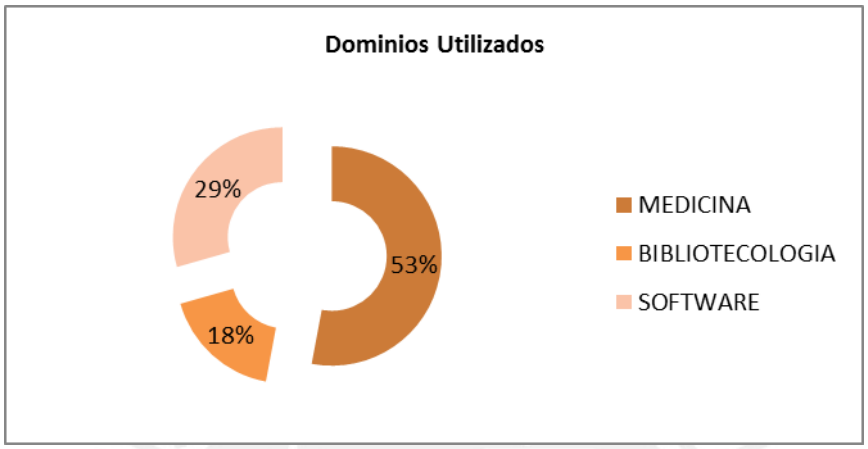


*Figura iv: Dominios Utilizados*

- **Soluciones Planteadas:** como ya se vio en la revisión del estado del arte, se analizaron una variedad de aplicaciones y soluciones que soportan la recuperación de conocimiento, usando o no ontologías, además estas aplicadas a diferentes dominios. Utilizar las ontologías como herramientas para la representación del conocimiento expresado en lenguajes globales como el RDF, son el común denominador en los artículos revisados. Se han podido extraer un conjunto de factores que hacen que los estudios realizados cumplan sus objetivos, estos son los siguientes:

    - Cumplir con algunos parámetros que la W3C recomienda, entre ellas el uso de RDF como lenguaje de etiquetación.
    - La delimitación de un dominio en particular, centrándose en las propiedades del mismo.
    - Recuperar conocimiento inferido utilizando ontologías.
    - Apoyar la idea de web semántica como un estándar en un futuro.

## 1.2 Consideraciones en el proyecto

En este anexo se han presentado los resultados de realizar una revisión sistemática sobre recuperación de conocimiento utilizando ontologías, que nos permite tener una

visión completa de la situación actual. Con la ayuda del protocolo de revisión aseguramos que nuestra investigación cumple con las características esenciales de una revisión sistemática propuesta por Schreiber. Este tipo de revision son un tanto mas complicadas de realizar a diferencia de las revisiones tradicionales.

Para sembrar una cultura orientada a cumplir los estatutos que propone la W3C con respecto a web semántica y así facilitar la recuperación de conocimiento, se debería comprometer a mas instituciones a manera que destinen inversiones en investigación y así administrar y crear conocimiento que sirva como *source* para futuras investigaciones.

Otro dato relevante obtenido es que existen un sin número de aplicaciones en la que la estructuración de la data para su posterior recuperación utilizando ontologías, mejoran considerablemente la eficacia de las búsquedas. Esto desarrollado en un dominio específico.

A partir de esta revisión sistemática ya se tiene conocimiento de la situación actual de los temas a tratar y se puede enfatizar en la realización de un modelo de recuperación de conocimiento utilizando lo más resaltante de los artículos revisados y aplicándolo al dominio de la Ingeniería Informática.