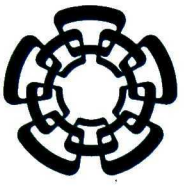


Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Laboratorio de Tecnologías de Información,
CINVESTAV-Tamaulipas

Obtención de Topic Maps a partir de Bases de Datos Relacionales

Tesis que presenta:

Adán José García

Para obtener el grado de:

Maestro en Ciencias en Computación

Director de la Tesis:
Dr. Iván López Arévalo

Cd. Victoria, Tamaulipas, México — Octubre, 2012

La tesis presentada por Adán José García fue aprobada por:

Dr. Javier Rubio Loyola

Dr. José Guadalupe Rodríguez García

Dr. Iván López Arévalo, Director

Cd. Victoria, Tamaulipas, México, 29 de Octubre de 2012

Agradecimientos

- Dedico este trabajo a mi familia, por acompañarme y apoyarme en cada decisión, por que son ustedes mi principal motivación. Especialmente a mi padre y amigo Roberto José José.
- Mi especial reconocimiento, admiración y respeto al Dr. Iván López Arévalo. Gracias por ser mi guía, por su paciencia y consejos que hicieron posible la culminación de este proyecto de investigación. Agradezco también a los revisores, el Dr. Javier Rubio, Dr. José Guadalupe Rodríguez y a la M.C. Ana Ríos Alvarado, por sus correcciones y criticas constructivas.
- Agradezco a mis amigos y compañeros del CINVESTAV, son muchos los momentos y anécdotas que compartimos, una experiencia grata de recordar. Especialmente a mis compañeros de maestría por hacer del CINVESTAV nuestro hogar.
- Gracias al cuerpo académico del Cinvestav, en especial al grupo de profesores que contribuyeron a mi formación profesional mediante enseñanzas y experiencias. Me siento orgulloso de formar parte del CINVESTAV.
- Extiendo un reconocimiento al CONACyT por incentivar e impulsar el desarrollo tecnológico en el país mediante apoyos económicos. ¡Gracias CONACyT!
- Agradezco a Mary por su constante apoyo, comprensión y cariño que me impulsaron a finalizar esta etapa. Este trabajo también te pertenece. ¡Mary, vamos por otra!
- A *DIOS*, por el regalo de la vida, por rodearme de seres extraordinarios, por mantener en mi esas ganas inmensas de luchar por cada meta trazada, por regalarme una vida llena de bendiciones.

Índice General

Índice General I

Índice de Figuras III

Índice de Tablas V

Índice de Algoritmos VII

Resumen IX

Abstract XI

Nomenclatura XIII

1. Introducción 1
- 1.1. Planteamiento del problema . . . 2
- 1.2. Objetivos generales y específicos . . . 3
  - 1.2.1. Objetivo principal . . . 3
  - 1.2.2. Objetivos específicos y contribuciones . . . 3
- 1.3. Estructura de la tesis . . . 5

2. Estado del Arte 7
- 2.1. Bases de datos . . . 8
  - 2.1.1. Modelo de datos relacional . . . 9
  - 2.1.2. Base de datos relacional . . . 11
- 2.2. Topic Maps . . . 13
  - 2.2.1. Modelo de datos de Topic Maps . . . 14
    - 2.2.1.1. Ejemplificación de tópicos, ocurrencias y asociaciones . . . 16
  - 2.2.2. Herramientas para Topic Maps . . . 17
- 2.3. Trabajos relacionados . . . 19
  - 2.3.1. Bases de Datos a Topic Maps . . . 20
  - 2.3.2. Integración de datos a partir de la generación de TMs . . . 22
  - 2.3.3. Obtención de TMs a partir de BDRs . . . 22
  - 2.3.4. Diseño de TMs a partir de recursos de información . . . 24

3. Metodología 29
- 3.1. Transformación de datos relacionales . . . 31
  - 3.1.1. Entradas al enfoque propuesto . . . 32
  - 3.1.2. Generación de tipos de tópicos . . . 38
  - 3.1.3. Generación de instancias de tipos de tópicos . . . 43

4. Implementación 49
4.1. Implementación del Enfoque Propuesto . . . 49
4.2. Integración de MOSTO . . . 53
4.2.1. Capa de presentación . . . 55
4.2.2. Capa de lógica . . . 55
4.2.3. Capa de datos . . . 56

5. Resultados 57
5.1. Banco de bases de datos relacionales . . . 57
5.1.1. Infraestructura de prueba . . . 60
5.2. Generación de Topic Maps . . . 60
5.3. Comparativa del volumen de datos . . . 64
5.3.1. Comparativa del volumen de información . . . 65
5.3.2. Análisis de resultados . . . 67
5.4. Validación sintáctica de Topic Maps . . . 69
5.4.1. Visualización de asociaciones . . . 70
5.4.2. Visualización con Vizigator . . . 72
5.5. Validación semántica de Topic Maps . . . 74
5.5.1. Esquema de validación semántica . . . 75
5.5.2. Reglas de inferencia . . . 79
5.5.3. Inferencia sobre el Topic Map *NORTHWIND* . . . 80

6. Conclusiones y Trabajo Futuro 83
6.1. Conclusiones . . . 83
6.2. Aportaciones . . . 85
6.3. Dificultades . . . 85
6.4. Trabajo Futuro . . . 87

Bibliografía 89

Índice de Figuras

2.1. Atributos y tuplas de la relación ORDERS 9
2.2. EBD de la BDR SALES 12
2.3. IBD de la BDR SALES 12
2.4. Norma ISO/IEC 13250 Topic Maps 13
2.5. Enfoque propuesto por Neidhart *et al.* [Neidhart et al., 2009] 21
2.6. Enfoque propuesto por Fei *et al.* [Ye et al., 2011] 23

3.1. Esquema general para la obtención de TMs a partir de datos relacionales 30
3.2. Interacción de las reglas de aprendizaje en el proceso de transformación . . . 31
3.3. Visualización de los conjuntos A, P, F y O para un esquema de relación R 32
3.4. EBD y RIs de referencia de la BDR SALES 33
3.5. IBD para el EBD SALES 34
3.6. Nombre de los elementos que conforman un tipo de asociación 35
3.7. Representación gráfica de los tipos asociaciones propuestos 35
3.8. Representación gráfica del tipo de asociación *es-atributo-de* 36
3.9. Representación gráfica del tipo de asociación *es-instancia-de* 36
3.10. Representación gráfica del tipo de asociación *tiene-relación-con* 37
3.11. Representación gráfica del tipo de asociación *tiene-referencia-con* 38
3.12. Variantes de los conjuntos P y F en una lista de atributos A 39
3.13. Plantilla para el modelado de los tipos de asociación (ejemplo: *es-atributo-de*) . . . 41

4.1. Diagrama de paquetes para la obtención de un TM 50
4.2. Diagrama de secuencia para la obtención de un TM 52
4.3. Arquitectura por capas para la integración del modelo propuesto MOSTO 54

5.1. Esquema general para la obtención automática de TMs 61
5.2. Comparativa de número de elementos TAOs 62
5.3. Comparativa del tiempo de cómputo porcentual para la generación de elementos TAOs 63
5.4. Comparativa del tiempo de cómputo porcentual promedio para la generación de los elementos TAOs 63
5.5. Comparativa del volumen de datos de entrada (BDR) y de salida (TM) 64
5.6. Comparativa del tiempo de cómputo entre el enfoque propuesto MOSTO y el propuesto por Eslami *et al.* 66
5.7. Comparativa del volumen de representación (elementos TAOs) entre el enfoque propuesto MOSTO y el propuesto por Eslami *et al.* 66
5.8. Comparativa del volumen de datos (KBs) entre el enfoque propuesto MOSTO y el propuesto por Eslami *et al.* 67
5.9. Visualización de las instancias del tipo de asociación *es-atributo-de* (is-attribute-of) . 70
5.10. Visualización de las instancias del tipo de asociación *es-instancia-de* (is-instance-of) 71

5.11. Visualización de las instancias del tipo de asociación *tiene-relacion-con* (has-a-relation-to) 71
5.12. Visualización de las instancias del tipo de asociación *tiene-referencia-con* (has-a-reference-to) 72
5.13. Visualización del TM SALES con la herramienta Vizigator 73
5.14. Tipos de predicados en Tolog 75
5.15. Componentes de una consulta básica con Tolog 75

Índice de Tablas

2.1. Lenguajes de consultas para Topic Map 18
2.2. Resumen de los enfoques propuestos para la construcción de TMs 27

3.1. Definición de los tipos de asociaciones y roles 35

5.1. Descripción del *benchmark* de BDRs propuesto para el proceso de experimentación . 58
5.2. Descripción de los TMs resultantes al aplicar el enfoque propuesto MOSTO . . . 61
5.3. Descripción de los TMs resultantes a partir del *benchmark* de BDRs 65
5.4. Comparativa de consultas entre SQL y Tologa para la instancia NORTHWIND 81

Índice de Algoritmos

1. **Identificación de tipos de ocurrencias en BDRs** 42

Resumen

Obtención de Topic Maps a partir de Bases de Datos Relacionales

por

Adán José García

Laboratorio de Tecnologías de Información, CINVESTAV-Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2012
Dr. Iván López Arévalo, Director

Las Bases de Datos Relacionales (BDRs) han sido tradicionalmente utilizadas como almacenamiento de datos por diversos sistemas de información. Considerando que las BDRs contienen datos valiosos y en gran escala, el desafío es mejorar la forma de acceder y compartir el conocimiento que reside en tales bases de datos. El uso de Topic Maps (TMs) es una solución para la organización y reutilización de dicho conocimiento. Sin embargo, el diseño manual de TMs es una tarea difícil que demanda tiempo, además de ser subjetiva si no existe una guía en común para su diseño. Los enfoques existentes sobre el diseño de los TMs no consideran el conocimiento que reside en las BDRs durante el proceso de transformación y, por lo tanto, el TM resultante no es una representación válida del sistema real. En este trabajo se presenta una metodología para la obtención automática de TMs a partir de BDRs. La metodología se basa en reglas de aprendizaje, las cuales guían el proceso de transformación y ayudan a capturar el conocimiento inherente en los datos relacionales. Para ello el proceso de transformación requiere del esquema, las instancias y las restricciones de integridad de la BDR. El proceso de extracción, manipulación y validación de los datos se realizó mediante expresiones regulares. La metodología propuesta fue implementada en Java e integrada a una aplicación web para la gestión de TMs. Se propuso un benchmark representativo de 15 BDRs para el proceso de experimentación. Los TMs resultantes fueron validados de manera sintáctica usando la herramienta Ontopia Vizigator y validados de manera semántica a través de reglas de inferencia empleando el lenguaje de consultas Tolog. Además, se presenta un análisis sobre el volumen de datos requerido para la representación del conocimiento de los datos relacionales.

Abstract

Building Topic Maps from Relational Databases

by

Adán José García

Information Technology Laboratory, CINVESTAV-Tamaulipas
Research Center for Advanced Study from the National Polytechnic Institute, 2012
Dr. Iván López Arévalo, Advisor

Relational Databases (RDBs) have been traditionally used as the backend database for information systems. Considering that RDBs contain valuable and massive data, the challenge is to find out how to improve accessing and sharing knowledge that resides in databases. The use of Topic Maps (TMs) is one solution for representing and reusing that knowledge. However, manual development of TMs is a difficult, time consuming and subjective task if there is not a common guideline. The existing TMs building approaches convert RDBs without considering the knowledge residing in the database during the transformation process and, therefore, the resulting TM is not a valid representation of the real system. This document proposes a methodology for automatically developing of TMs from BDRs. The proposed methodology is based on learning rules, which are a guideline for the transformation process and help to capture the inherent knowledge in the relational data. Then, the transformation process requires the schema database, the instances and the integrity constraints of RDB. The extraction, manipulation and data validation process was done by regular expressions. The proposed approach was implemented in Java and was integrated to a web application for the generation and management of TMs. A representative benchmark of 15 RDB was proposed for the experimentation process. The resulting topic maps were validated syntactically using the Ontopia Vizigator tool and validated semantically through the inference of information using the Tolog query language. In addition, we performed a study about the volume data required for the knowledge representation of the relational data. Finally, the results found in our experimentation are encouraging and they show the soundness of the proposed methodology.

Nomenclatura

RC	Representación del Conocimiento
WS	Web Semántica
TM	Topic Map
MDTM	Modelo de Datos de Topic Maps
LCTM	Lenguaje de Consultas para Topic Maps
NGTM	Notación Gráfica para Topic Maps
XTM	XML for Topic Maps
TAO	Tópico + Asociación + Ocurrencia
BD	Base de Datos
MD	Modelo de Datos
EBD	Esquema de la Base de Datos
IBD	Estado de la Base de Datos
SMBD	Sistema Manejador de Base de Datos
BDR	Base de Datos Relacional
MDR	Modelo de Datos Relacional
RI	Restricción de Integridad
ER	Expresión Regular

1

Introducción

En este capítulo se describe brevemente el planteamiento del problema, los objetivos y las principales contribuciones de este trabajo de investigación.

Las ontologías son un campo de investigación de la Inteligencia Artificial y más específicamente de la rama relacionada con la *Representación del Conocimiento* (RC), cuyo objetivo principal es facilitar el intercambio y la reutilización de conocimientos [Davis et al., 1993].

En este contexto, la *Web Semántica* (WS) es el principal impulsor de la RC, la cual surge de las propuestas de Tim Berners-Lee [Lee et al., 2001], que apuntaban hacia el enlazado de los conceptos y hechos a una escala global. Esta idea inicial de la Web se ha difundido con el calificativo de Web Semántica, "una extensión de la Web actual en la que la información tiene un significado bien definido". Donde, el objetivo es llevar la Web a su máximo potencial, es decir, al máximo nivel de automatización en los procesos de transferencia de la información y el conocimiento.

En este sentido, los *Topic Maps* (TMs) son una tecnología estándar de la WS, para la anotación semántica de recursos de información [Garshol, 2002]. Los TMs se utilizan para organizar la información, así como para gestionarla y facilitar su recuperación.

Existen muchos recursos de información que pueden modelarse mediante TMs. En este sentido, las *Bases de Datos Relacionales* (BDRs) representan un recurso de información que puede ser aprovechado de múltiples formas al obtener su representación semántica. Según Elmasri y Navathe [Elmasri and Navathe, 2011], el *Modelo de Datos Relacional* (MDR) es el más utilizado en la actualidad para la gestión de bases de datos por distintos sistemas de información. Así, las BDRs representan el medio de almacenamiento de datos más utilizado por la industria, el gobierno y el ámbito educativo.

Por lo tanto, los TMs pueden emplearse para describir y representar un área especifica de conocimiento (por ejemplo una BDR). Además de expresar las relaciones entre los datos por medio de un lenguaje formal (lógico) que puede ser entendido por una computadora.

1.1 Planteamiento del problema

Las *Bases de Datos Relacionales* (BDRs) se han utilizado ampliamente como principal medio de almacenamiento de datos de los sistemas de información [Elmasri and Navathe, 2011]. Considerando que las BDRs contienen información valiosa en gran escala, el desafío es encontrar una manera de representar y compartir el conocimiento que reside en ellas. En este contexto la WS, mediante la propuesta de nuevos paradigmas y tecnologías web, propone tener datos enlazados con un significado bien definido, ofreciendo una plataforma accesible que permita que los datos se compartan y se procesen por herramientas automatizadas [Lee et al., 2001].

Los Topic Maps (TMs) constituyen una de las técnicas de la WS para brindar significado a la información. Los TMs proveen una gramática estándar para la anotación semántica de recursos de información dentro de un dominio especifico [Garshol, 2002]. Los TMs no solo ayudan a almacenar, recuperar y modificar grandes cantidades de información como las BDRs, sino también, permiten plasmar elementos de conocimiento. No obstante, el diseño manual de TMs es una tarea difícil que demanda tiempo, además de ser subjetiva sino se cuenta con una guía base. Por lo tanto, la obtención de un Topic Map es un tarea difícil, donde se requiere de un buen diseño para asegurar que el resultado refleje el sistema de información real que se representa. Por buen diseño nos referiremos a una

representación de dominio que guarde relaciones coherentes entre los datos y que ésta representación pueda ser validada semánticamente.

En este trabajo de tesis se propone una metodología basada en reglas para la obtención automática de TMs a partir del esquema y los datos de BDRs. Los TMs resultantes se validan de forma sintáctica mediante una herramienta estándar reportada en la literatura especializada. Además, se propone un esquema de validación semántica de TMs mediante reglas de inferencia usando un lenguaje de consultas estándar para TMs.

1.2 Objetivos generales y específicos

Este trabajo de tesis aborda el diseño de una metodología para la representación del conocimiento de las BDRs mediante la construcción automática de TMs válidos semánticamente. Los enfoques propuestos anteriormente requieren de una configuración previa que les indique cómo deben modelarse los datos relacionales. A continuación se describen los objetivos que persigue el presente trabajo de investigación.

1.2.1 Objetivo principal

El objetivo principal de esta investigación es contribuir en el avance del estado del arte en el campo de la Representación del Conocimiento con una metodología basada en reglas para la obtención automática de TMs a partir de BDRs.

1.2.2 Objetivos específicos y contribuciones

Con el fin de cumplir con el objetivo principal, este trabajo de tesis contempla tres objetivos específicos. Las contribuciones de este trabajo son los resultados directos del cumplimiento de cada objetivo específico. Es preciso señalar que los objetivos específicos que se describen a continuación no reflejan un orden de relevancia:

- **Obtener un conjunto de reglas de aprendizaje que permitan representar el conocimiento extraído de las** BDRs. Esta primer contribución consiste en proponer un conjunto de reglas de aprendizaje cuyo objetivo es guiar el proceso de transformación a partir de datos relacionales y obtener una representación semántica de los mismos. Durante el proceso de transformación, todos los elementos semánticos que conforman el TM resultante (tópicos, asociaciones y ocurrencias) se obtienen mediante algoritmos que emplean expresiones regulares para su identificación, es por eso que se considera un proceso automático. Además, se propone un conjunto de tipos de asociaciones necesarias para representar los enlaces semánticos entre los recursos de información. De acuerdo a los resultados experimentales obtenidos, el enfoque propuesto basado en reglas fue capaz de obtener y representar el conocimiento extraído de las BDRs.

- **Definir un esquema de validación sobre el conjunto de reglas de aprendizaje propuestas**. Establecer las condiciones necesarias para la validación semántica entre los elementos de los TMs mediante la navegabilidad e inferencia, empleando reglas de inferencia para la recuperación de información, gestión del conocimiento y mantenimiento de los TMs.

- **Definir un *benchmark* de prueba para la validación de resultados**. Definición de un conjunto de BDRs representativas provenientes de distintos *Sistemas Manejadores de Base de Datos* (SMBDs) (MySQL, SQL Server® y Oracle®). Este *benchmark* posee las características necesarias y suficientes para la prueba del enfoque semántico propuesto. Algunas de estas características que posen las BDRs que integran el benchmark son: a) fueron construidas a partir de información real y ficticia, b) poseen un número distinto de relaciones, atributos e instancias, c) poseen múltiples restricciones de integridad de referencia, d) poseen valores nulos y campos vacíos y finalmente e) poseen redundancia de datos.

1.3 Estructura de la tesis

Esta tesis está organizada en 6 capítulos. Los 2 primeros capítulos describen los conceptos de fondo necesarios para comprender el resto del contenido de la tesis. Los 4 últimos capítulos presentan las contribuciones, sus resultados correspondientes y las conclusiones de la tesis:

- En el Capítulo 2 se presentan los conceptos básicos que se requieren para entender el resto del documento. Además, se presenta una revisión del estudio del estado del arte de los enfoques propuestos para el diseño de TMs a partir de BDRs.
- En el Capítulo 3 se describe el enfoque propuesto basado en reglas para el diseño automático de TMs a partir del esquema e instancias de la BDR. Además se presenta un algoritmo para la identificación de tipos de ocurrencias.
- En el Capítulo 4 se presenta la implementación del enfoque propuesto y su integración al sistema de prueba de concepto.
- En el Capítulo 5 se presenta los mecanismos de evaluación del enfoque propuesto. Se presenta una comparativa con el enfoque más representativo del estado del arte. Finalmente se presentan los resultados obtenidos de la visualización y validación semántica de TMs.
- En el Capítulo 6 se proporcionan algunos comentarios finales y posibles trabajos de investigación futura.

2

Estado del Arte

En este capítulo se presentan los conceptos básicos relacionados con los Topic Maps (TMs) *y las Bases de Datos Relacionales* (BDRs)*, además se presenta un estudio del estado del arte sobre el diseño de* TMs *a partir de datos relacionales.*

La Web surge como una manera de compartir información, en forma de documentos estáticos [Davis et al., 1993]. Sin embargo, desde su creación hasta la actualidad, la Web sigue ofreciendo nuevas posibilidades y usos no previstos inicialmente. Actualmente la Web es un gran almacén de datos, donde su potencial como gestor de conocimiento universal se ve acrecentado exponencialmente cuando se le otorga un recubrimiento semántico a esos mismos datos, abriendo un mayor espectro de interoperabilidad entre aplicaciones web [Mark, 2007]. Esta extensión de la Web, conocida como *Web Semántica* (WS), se beneficia de distintos factores como el desarrollo, aceptación y éxito del lenguaje de marcas XML (*eXtensible Markup Languaje)* y del uso de ontologías para la descripción del conocimiento de un dominio específico, como por ejemplo el dominio semántico de las BDRs. En los siguientes apartados se describen los conceptos básicos relacionados con las BDRs y los conceptos relacionados a la tecnología estándar TMs para la representación de los mismos.

2.1 Bases de datos

Las Bases de Datos (BDs) desempeñan un papel crítico en la mayoría de las áreas donde se utilizan las computadoras [Gilleson, 2005]. Tradicionalmente, la tecnología de BDs se aplica a datos estructurados que surgen de los procesos de rutina en el gobierno, los negocios y la industria. Una BD puede verse como una colección de datos relacionados. Por datos nos referimos a hechos conocidos que pueden registrarse y que tienen un significado implícito [Elmasri and Navathe, 2011]. La definición anterior de BDs es bastante general, sin embargo el término *base de datos* tiene las siguientes propiedades implícitas:

- Representa aspectos del mundo real, conocido como el universo de discurso.
- Se concibe como una colección lógicamente coherente de datos con un significado inherente.
- Se ha diseñado, construido y poblado con datos para un propósito específico.

Para la implementación y mantenimiento de una BD es necesario un *Sistema Manejador de Base de Datos* (SMBD). El SMBD es un sistema de software de propósito general que facilita los procesos de definición, construcción, manipulación e intercambio de BDs entre distintos usuarios y aplicaciones.

Una característica fundamental del enfoque de una BD es que proporciona cierto nivel de abstracción de datos. La abstracción de los datos generalmente se refiere a la supresión de los detalles de organización y almacenamiento de datos. Un *Modelo de Datos* (MD) es una colección de conceptos que se utilizan para describir la estructura de una BD y proporciona los medios necesarios para lograr esta abstracción. Por estructura de una BD nos referimos a los tipos de datos, relaciones y restricciones que se aplican a los datos.

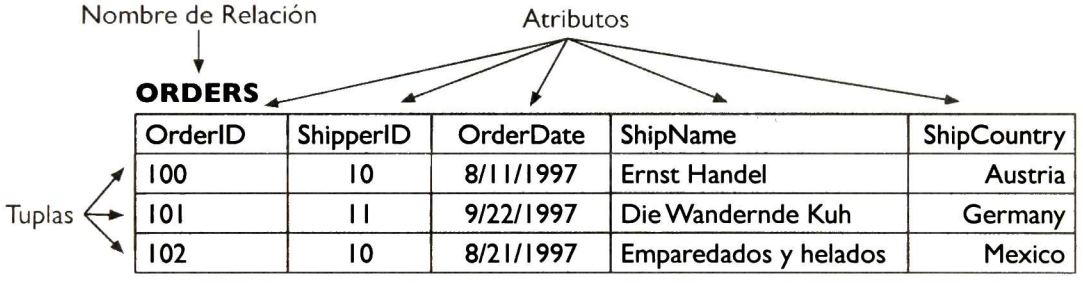
En cualquier MD es importante distinguir entre la descripción de la BD y de la propia base de datos. En este trabajo de tesis, a la descripción de una BD se le denomina *Esquema de la Base de Datos* (EBD), la cual se especifica durante el diseño de la base de datos. Por otro lado, a los datos o instancias de la BD se les denomina *Estado de la Base de Datos* (IBD).

2.1.1 Modelo de datos relacional

El *Modelo de Datos Relacional* (MDR) utiliza el concepto de una relación matemática como su componente básico, y tiene su base teórica en la teoría de conjuntos y la lógica de predicados de primer orden [Elmasri and Navathe, 2011]. El MDR representa una BD como una colección de *relaciones*.

Cuando una relación se considera como una tabla de valores, cada fila de la tabla representa una colección de valores de datos relacionados. Una fila representa un hecho que normalmente corresponde a una entidad del mundo real. El nombre de la tabla y los nombres de las columnas se utilizan para ayudar a interpretar el significado de los valores en cada fila.

En la terminología formal del MDR, una fila se denomina *tupla*, la cabecera de una columna se denomina *atributo* y la tabla se denomina *relación* (ver Figura 2.1). El tipo de dato que describe los tipos de valores que pueden aparecer en cada columna se representa por un *dominio* de posibles valores. A continuación se definen estos términos de manera formal:



OrderID	ShipperID	OrderDate	ShipName	ShipCountry
100	10	8/11/1997	Ernst Handel	Austria
101	11	9/22/1997	Die Wandernde Kuh	Germany
102	10	8/21/1997	Emparedados y helados	Mexico

Figura 2.1: Atributos y tuplas de la relación ORDERS

Un **dominio** D es un conjunto de valores atómicos. Por atómico nos referimos a que cada valor en el dominio es indivisible hasta donde el modelo relacional se refiera. Un método común para la especificación de un dominio consiste en determinar qué valores forman el domino para un tipo de dato dado.

Definición 1 Un **esquema de relación** R, denotado por $R(A_1, A_2, \ldots, A_n)$, se compone de un nombre de relación R y una lista de atributos, $A_1, A_2, ..., A_n$. Cada atributo A_i es el nombre de un rol desempeñado por algún dominio D en el esquema de relación R, y es denotado por

$dom(A_i)$. El *grado* (o aridad) de una relación es el número de atributos n de su esquema de relación.

De acuerdo a la Figura 2.1, el estado de relación ORDERS es denotado como ORDERS(OrderID, ShiperID, OrderDate, ShipName, ShipCountry). Donde el atributo ShipCountry desempeña el rol de *país* de envío de paquetes, donde su dominio es una secuencia de caracteres que forman el nombre de un país válido. Por otra parte, de acuerdo al número de atributos que integran la relación ORDERS, se dice que la relación es de grado 5.

Definición 2 Una **relación** (o *estado de relación* [1]) r de un esquema de relación $R(A_1, A_2, \ldots, A_n)$, también denotado por r(R), es un conjunto de *n*-tuplas $r = \{t_1, t_2, \ldots, t_p\}$. Cada *n*-tupla t es una lista ordenada de n valores $t = \langle v_1, v_2, \ldots, v_n \rangle$, donde cada valor v_i, $1 \leq i \leq n$, es un elemento de $dom(A_i)$. El *i*-ésimo valor en la tupla t, que corresponde al atributo A_i, es referenciado como $t[A_i]$.

La Figura 2.1 representan el estado de la relación ORDERS. El estado de la relación está formada por un conjunto de 3 tuplas. La tupla $t_1 = \langle 100, 10, 8/11/1997, \text{Ernst Handel}, \text{Austria} \rangle$ es una lista de 5 valores, donde el valor *"Austria"* que corresponde al atributo ShipCountry y es referenciado como $t_1[A_5]$.

A continuación se presenta una definición formal de *relación* empleando conceptos de la teoría de conjuntos:

Una relación matemática r(R) de grado n en los dominios $dom(A_1), dom(A_2), \ldots, dom(A_n)$ es un subconjunto del producto Cartesiano (denotado por $\times$) de los dominios que definen R:

$$r(R) \subseteq (dom(A_1), dom(A_2), \ldots, dom(A_n))$$

El producto cartesiano especifica todas las combinaciones posibles de valores de los dominios subyacentes. Por lo tanto, si denotamos el número total de valores, o cardinalidad, en un dominio D por $|D|$, el número total de tuplas en producto cartesiano es

[1] También se le llama relación de instancias. En este documento no se utilizará este término porque instancia es también usado para referirse a una sola tupla.

$$|\text{dom}(A_1)| \times |\text{dom}(A_2)| \times \cdots \times |\text{dom}(A_n)|$$

Este producto de las cardinalidades de todos los dominios representa el número total de instancias posibles o tuplas que siempre pueden existir en cualquier relación r(R).

2.1.2 Base de datos relacional

Las definiciones y restricciones definidas hasta el momento aplican únicamente a una relación y sus atributos. Una BDR usualmente contiene muchas relaciones, y las relaciones contienen múltiples tuplas que se relacionan de múltiples maneras. A continuación se define el *Esquema de la Base de Datos* (EBD) y el *Estado de la Base de Datos* (IBD):

Definición 3 Un **esquema de la base de datos relacional** S es un conjunto de *esquemas de relación* $S = \{R_1, R_2, \ldots, R_m\}$ y un conjunto de *Restricciones de Integridad* (RI). Mientras que el **estado de la base de datos relacional**[2] E obtenido a partir de S, es un conjunto de *estados de relación* $E = \{r_1, r_2, \ldots, r_m\}$, tal que cada r_i es un estado de R_i y tal que cada estado de relación r_i satisface a las restricciones de integridad especificadas en RI.

La Figura 2.2 muestra el EBD denominado SALES = {SHIPPERS, ORDERS, PRODUCTS, ORDER_DETAILS}. Los atributos subrayados representan llaves primarias. Mientras que, la Figura 2.3 muestra el IBD correspondiente al EBD SALES. La definición del EBD y el IBD de SALES se utilizará en el Capítulo 3 y 5 como ejemplo para la demostración del enfoque propuesto.

En este trabajo de investigación cuando se haga mención a una *Base de Datos Relacional* (BDR) implícitamente se hace referencia tanto a su *Esquema de la Base de Datos* (EBD) denotado por S como a su *Estado de la Base de Datos* (IBD) denotado por E.

[2]El EBD es también denominada como instancia de la base de datos.

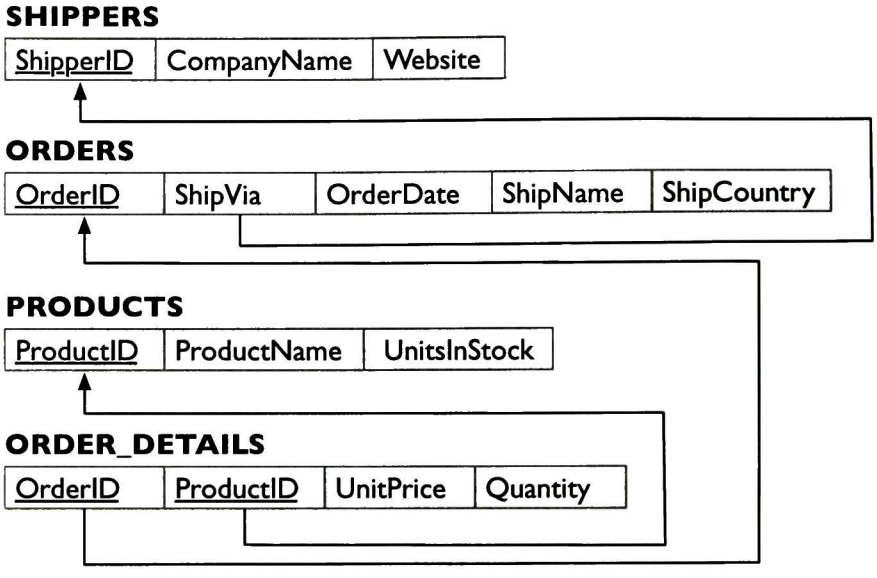


Figura 2.2: EBD de la BDR SALES

ORDERS

OrderID	ShipVia	OrderDate	ShipName	ShipCountry
100	10	8/11/1997	Ernst Handel	NULL
101	11	9/22/1997	Die Wandernde Kuh	Germany
102	10	8/21/1997	Emparedados y helados	Mexico

ORDER DETAILS

OrderID	ProductID	UnitPrice	Quantity
100	51	45.90	12
101	50	99.34	55
102	51	45.90	30

PRDUCTS

ProductID	ProductName	UnitsInStock
50	Chocolate	200
51	Filo Mix	23

SHIPPERS

ShipperID	CompanyName	Website
10	Speed Express	http://www.speedexpress.com
11	United Package	http://www.unitedpackage.com

Figura 2.3: IBD de la BDR SALES

2.2 Topic Maps

Los TMs tienen su origen en el grupo Davenport[3], como una norma para la fusión de índices impresos. Posteriormente evolucionó hacia otras estructuras, hasta llegar a ser considerada una herramienta web para la organización, representación y gestión del conocimiento.

La primera versión oficial del estándar ISO/IEC 13250:Topic Maps data del año 2000 [ISO/IEC, 1999]. La norma engloba un conjunto de estándares que definen el modelo y la sintaxis de intercambio para su formalización entre diferentes aplicaciones. La Figura 2.4 muestra el conjunto de normas que se han construido para la tecnología Topic Maps. Se observa la relación que existe entre el Modelo de Datos de Topic Maps (TMDM, por sus siglas en inglés) con los distintos estándares propuestos para la materialización de los TMs (CTM[4], XTM[5]). Se presenta el estándar para el Lenguaje de Consultas para Topic Maps (TMQL, por sus siglas en inglés) y además se presenta el estándar para la Graficación de Topic Maps (GTM, por sus siglas en inglés).

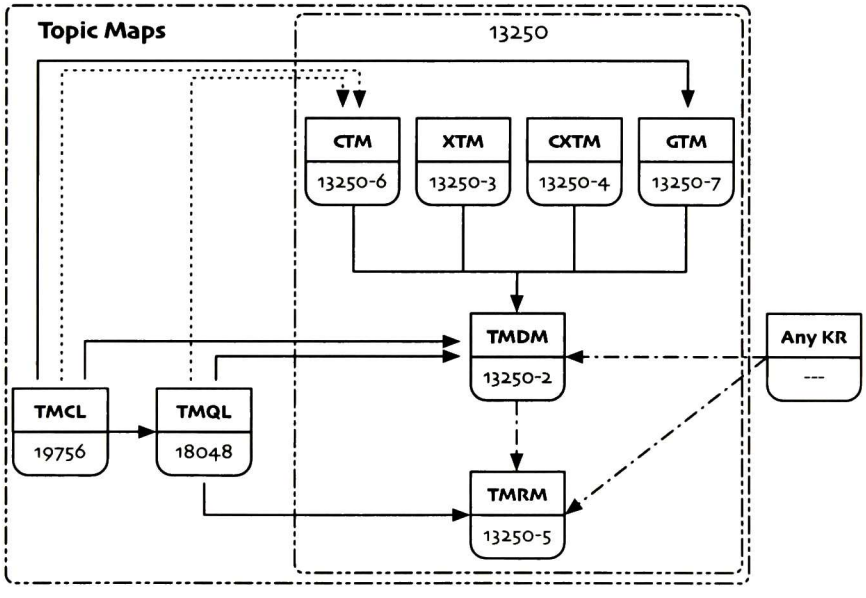


Figura 2.4: Norma ISO/IEC 13250 Topic Maps

[3]El Grupo de Davenport es un grupo de productores y editores de libros electrónicos.
[4]Compact Syntax for Topic Maps
[5]XML for Topic Maps

En un inicio el estándar TM se ideó para la arquitectura SGML[6] con notación de HyTime[7]. No obstante, esta notación cayó en desuso tras la creación de una Definición de Tipo de Documento (DTD) para crear Topic Maps en XML, denominada *XML for Topic Maps* (XTM). Dada la novedad del estándar Topic Map y su escasa implantación en la Web, se ha considerado oportuno revisar y definir sus elementos principales del modelo, analizando sus posibilidades para la organización del conocimiento, como estructura de navegación semántica. Nos centraremos entonces en el modelo descrito para la sintaxis XML, es decir, la especificación XTM en su versión 2.0 [ISO/IEC, 2006b], que es el lenguaje que está impulsando el desarrollo de la WS. Además, se ha optado por mantener los términos originales en inglés para evitar ambigüedades en la descripción del enfoque propuesto.

2.2.1 Modelo de datos de Topic Maps

El núcleo central del *Modelo de Datos de Topic Maps* (MDTM) definido por el estándar ISO/IEC 13250 Topic Maps [ISO/IEC, 2006a] está constituido por tres elementos básicos: *Topic*, *Association*, y *Occurrence*. Esta triada de conceptos se conoce como el TAO[8] de los *Topic Maps* (TMs), concepto anunciado por Steve Pepper [Pepper, 2002], uno de los editores de XTM. A continuación se describen los elementos básicos de los TMs.

Topic es el elemento principal de un TM, el cual constituye la representación material o concreta del *subject*, que es una percepción humana abstracta de una realidad. Entre *topic* y *subject* se establece una relación biunívoca en la cual un *subject* es representado por un único *topic* y viceversa. La norma ISO define a un *subject* como cualquier cosa, con independencia y características específicas, sobre la cual puede decirse cualquier cosa con cualquier significado.

Los *topics* tienen tres características principales: su denominación (*topic name*), sus descripciones (*occurrence*) y un rol como miembro de una asociación (*role association*). Esta asignación de características se considera válida para un determinado *scope* (contexto), donde dos *topics* con las mismas características se consideran idénticos.

[6]Estándar de Lenguaje de Marcado Generalizado para la organización y etiquetado de documentos.
[7]Lenguaje estructurado basado en hipermedia utilizado para la presentación y sincronización de contenidos.
[8]TAO = Topic + Association + Occurrence

Un *topic name* hace referencia a las diferentes formas de denominación que puede tener un *topic*. El modelo permite definir nombres normalizados a los *topics* que sean significativos desde el punto de vista semántico. Así, un topic puede tener varias denominaciones mediante la asignación de múltiples *base name*. Además, el *topic* puede tener otras denominaciones, como *display name*, que es la forma en la se mostrará al usuario y un *sort name*, que es como se ordenará alfabéticamente.

Cada *topic* es una instancia de una o más clases de *topics* (denominados también *topic type*), que pueden o no indicarse de forma explícita. Los ***topic type*** definen relaciones clase-instancia y la especificación XTM lo materializa mediante el elemento `instanceOf`.

Occurrence es cualquier información relevante para un *subject* dado. Representan recursos externos de información, que aclaran o ejemplifican el significado del topic. Los recursos pueden ser de muchos tipos, una cita textual, una definición, una fórmula, una URI, etc. Existen dos tipos de *occurrences*: *i)* `resourceRef`, que es un enlace a un recurso externo de información y *ii)* `resourceData`, que es algún dato interno que se facilita para la interpretación del tópico.

Cada *occurrence* es instancia de sólo una clase de *occurrence* (denominada también ***occurrence type***), que puede o no indicarse de forma explícita y se expresa mediante el elemento `instanceOf`.

Association representa una relación entre uno o más *topics*, donde cada uno de ellos juega un rol como miembro de dicha relación. El *topic* no se define solo por su denominación, sino también por sus relaciones y su contexto. El número de *topics* involucrados en una asociación no está limitado, aunque lo más frecuente es que sean dos (asociaciones binarias) o, en mayor grado, tres (asociaciones ternarias). Al igual que los *topics* y *occurrences*, las asociaciones se pueden agrupar en clases o ***association type***.

La capacidad de crear una gran diversidad de clases de agrupaciones que permiten establecer asociaciones entre *topics*, de acuerdo a las posibilidades que ofrece la lógica, permiten el desarrollo de tecnologías para navegar por conjuntos grandes de información.

2.2.1.1. *Ejemplificación de tópicos, ocurrencias y asociaciones*

Un tópico es el término que expresa determinado concepto o idea. Ejemplo de tópicos pueden ser *"Europa"*, *"persona"*, *"continente"*. Los tópicos pueden asociarse con otros tópicos denominados tipos de tópicos, p.e *"Europa"* puede tener un tipo de tópico que sea *"continente"*. Los tipos de tópicos definen relaciones *clase-instancia*, para otro tipo de relaciones se deben de crear asociaciones específicas.

Se mencionó que los tópicos tienen tres características principales. A continuación se describen estas características:

- Nombres: Un tópico puede tener varias denominaciones, pero debe estar presentado por una forma base *(base name)*. El nombre base es obligatorio y representa la forma usual de hacer mención al tópico de forma no ambigua.

- Ocurrencias: Son enlaces a recursos informativos y pueden ser de muchos tipos, p.e. una página web. Cada uno de los diferentes tipos puede ser agrupado mediante roles de ocurrencia (p.e. diccionario, página web, imagen, etc). Cuando se diseña un TM es preferible que el número de ocurrencias se limite a unos pocos recursos muy relevantes.

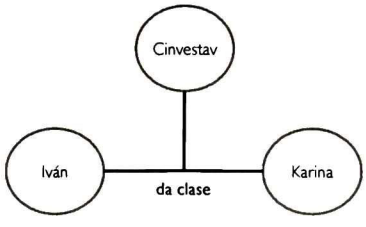
- Asociaciones: La asociación es un enlace que establece una relación entre dos o más tópicos. Una forma de ejemplificarlo es tomar una frase y considerar los sustantivos de esas frases como tópicos empleando los términos de unión entre los tópicos para denominar la asociación. Por ejemplo:

  - *"Tamaulipas es una entidad federativa de México"*
  - *"El cloruro de sodio se denomina sal"*
  - *"En Oaxaca encontramos diversidad cultural"*
  - *"La miel se localiza en los granos de café maduro"*
  - *"El Cinvestav se encuentra en Tamaulipas"*

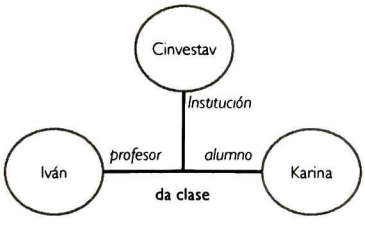
Las asociaciones se pueden agrupar por tipos. De este modo se puede agrupar *"se encuentra en"* y *"se localiza en"* como un tipo de asociación de *"ubicación"*.

Los roles en las asociaciones ayudan a disminuir la ambigüedad. Entendemos por tipo de rol al papel que desempeña un determinado tópico en una asociación. En la frase *"El Cinvestav se encuentra en Tamaulipas"* tenemos que *"Cinvestav"* desempeña un rol de institución y el término *"Tamaulipas"* designa un rol de lugar y representa un estado.

En una asociación, los tópicos se denominan miembros y los miembros representan roles, por ejemplo de la frase *"Iván da clase a Karina en el Cinvestav"* se puede obtener la siguiente representación ternaria:



Sin añadir más información no es posible saber quién es el profesor y quién el alumno. Esta información se puede expresar mediante roles:



Los roles *"alumno"*, *"lugar"*, *"profesor"* deberán ser definidos como tópicos, al igual que *"Iván"*, *"Karina"* y *"da clase"*.

2.2.2 Herramientas para Topic Maps

A continuación se describen algunas de las herramientas más representativas para visualizar y/o gestionar Topics Maps.

Tabla 2.1: Lenguajes de consultas para Topic Map

NOMBRE	DESCRIPCIÓN	AUTOR	ESTADO
Tolog [Garshol, 2006]	A Topic Maps Query Language	Ontopia	En producción
AsTMa [Barta, 2004]	Asymptotic Topic Map Notation, Constraining	Robert Barta	En producción
TMRQL [Moore and Ahmed, 2005]	Topic Map Relational Query Language	Graham Moore y Kal Ahmed	Prototipo
TOMA [Rani et al., 2007]	Topic Map Query Language, Topic Map Manipulation Language and Topic Map Constraint Language	Rani Pinchuk	Prototipo
TMSparql [Ahmed, 2009]	Making Topic Maps SPARQL	Kal Ahmed	Prototipo
XTMPath [Gylta and Barta, 2002]	Manipulating Topic Map Data Structures	Jan Gylta y Robert Barta	En producción

- Ontopia [Pepper, 2012] proporciona un conjunto de herramientas para la gestión del conocimiento como: Ontopia Topic Map Engine, Ontopia Navigator Framework y The Omnigator, entre otras. Todas ellas utilizan estándares que hacen posible el intercambio de información entre distintos sistemas.

- Wandora [Wandora, 2012] es un aplicación de propósito general para la extracción de datos, gestión y publicación de aplicaciones basado en Topic Maps y Java. Proporciona una interfaz gráfica de usuario, una capa de presentación del conocimiento, varias opciones de almacenamiento de datos, opciones de extracción de datos e integra un servidor HTTP para la publicación automática de Topic Maps.

- Un Lenguaje de Consultas para Topic Maps (LCTM) es un estándar ISO para consultar y manipular información de un Topic Map. La Tabla 2.1 describe los principales enfoques sobre lenguajes de consultas.

- Visualización de Topic Maps. La navegación permite a los usuario visualizar el Topic Map y

acceder a la información de manera rápida. La navegación debe ser intuitiva tal que permita desplazarse de un lugar a otro con facilidad.

- Vizigator [Ontopia, 2010]. Herramienta desarrollada por Ontopia, permite visualizar de manera gráfica la estructura de un Topic Map, la cual es útil para visualizar grandes y complejas cantidades de datos, o simplemente visualizar de forma atractiva un Topic Map.
- GTMalpha [Redmann et al., 2008]. Es una herramienta para la representación gráfica conceptual de un Topic Map.

2.3 Trabajos relacionados

Se han realizado numerosas investigaciones sobre el diseño de TMs a partir de distintos recursos de información. Existen tres principales enfoques para la creación de TMs:

- Construcción manual. Los TMs resultantes usualmente presentan una buena calidad de diseño y alto desempeño en aplicaciones particulares. Sin embargo, su diseño manual es una tarea costosa que demanda conocimientos técnicos y tiempo prolongado de elaboración cuando se trata de muchos recursos de información.
- Construcción automática a partir de datos no estructurados. Para la creación de los TMs es necesario el uso de herramientas de procesamiento de lenguaje natural.
- Construcción automática a partir de datos estructurados. Distintos trabajos de investigación se han propuesto para el diseño de TMs a partir de datos estructurados como XML, Hipertexto, BDRs y otras aplicaciones especializadas.

La publicación del modelo *XML for Topic Maps* (XTM) como estándar por la *International Standard Organization* (ISO) en el año 2000 y su posterior adaptación al lenguaje XML mediante la especificación XTM, ha despertado gran interés en la comunidad científica. En años recientes ha cobrado especial interés el diseño de TM a partir de datos relacionales, debido a las características

que posee el estándar *XML for Topic Maps* (XTM) para la representación del conocimiento aunado a la necesidad de integrar y compartir el conocimiento valioso que reside en las BDRs. Para ello existen dos técnicas principales:

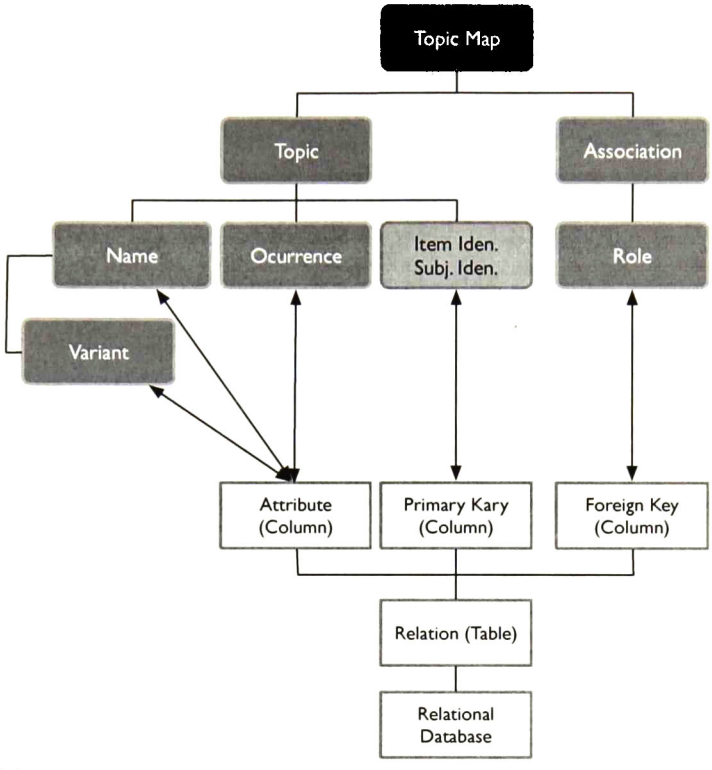
- basadas en ingeniería inversa [Ontopia, 2011, Neidhart et al., 2009, Ye et al., 2011] y
- basadas en el modelado del *Esquema de la Base de Datos* (EBD) [Eslami and Nazami, 2011].

2.3.1 Bases de Datos a Topic Maps

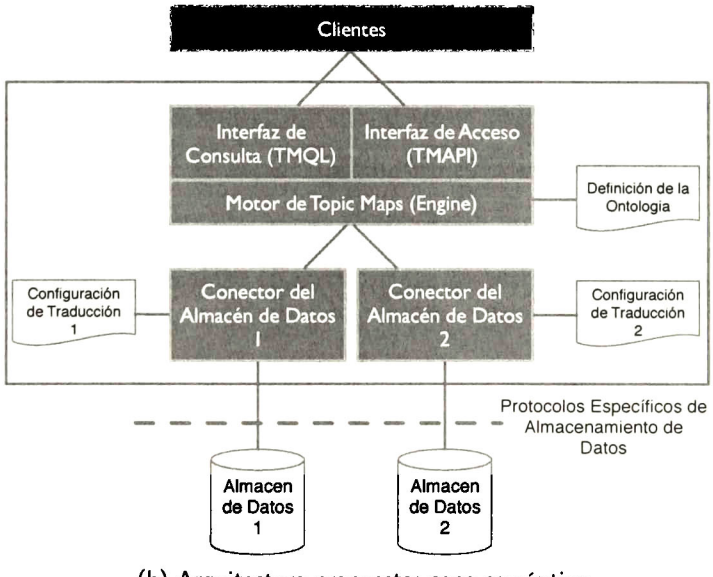
El módulo DB2TM [Ontopia, 2011] requiere de un TM (probablemente vacío), un archivo de configuración en formato XML donde se especifica la asignación de BDR a la ontología TM (conjunto de declaraciones de relaciones) y la fuente de datos relacional (el directorio que contiene la BDR). En [Neidhart et al., 2009] se presenta el prototipo de una capa de integración semántica para la disposición transparente de datos relacionales a través del uso de TMs. Emplean la tecnología TM para la representación de los datos, además se incorpora información de metadatos, permitiendo así la integración a nivel semántico. El enfoque propuesto define un conjunto de lineamientos para la asignación de la estructura de datos relacionales:

- Las llaves primarias se modelan como identificadores de recursos (o *subject identificator*).
- Los nombres de las columnas (o atributos) se modelan como tipos de ocurrencias o tipos de tópicos.
- Las llaves foráneas se modelan como tipo de roles en asociaciones.

La Figura 2.5a presenta el enfoque propuesto por Neidhart *et al.* para la asignación de la BDR a la ontología TMs. Por otro lado, la Figura 2.5b presenta la arquitectura completa de la capa de integración semántica propuesta. Además, la capa se integra al motor de TM *TopiEngine* para la gestión y almacenamiento del TM.



(a) Modelado entre las estructuras de datos relacionales y su conversión a TM



(b) Arquitectura propuesta: capa semántica

Figura 2.5: Enfoque propuesto por Neidhart *et al.* [Neidhart et al., 2009]

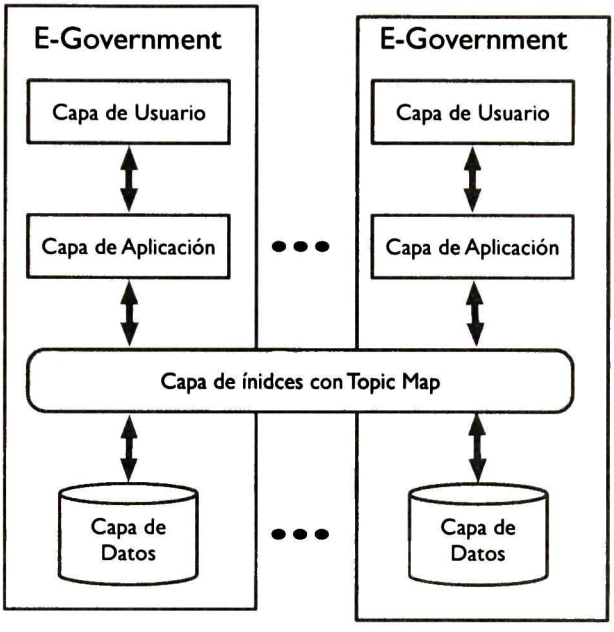
2.3.2 Integración de datos a partir de la generación de TMs

En [Ye et al., 2011] se presenta un análisis de BDRs heterogéneas de distintos sistemas de gobierno. Los autores presentan el prototipo de un modelo de integración de datos basado en la generación y combinación automática de TMs. La Figura 2.6a presenta la capa de índices de TMs encargada de la integración de datos entre la capa de aplicación y la capa de datos. La capa de índice extrae datos del SMBD y emplea la tecnología TM para constituir una indexación semántica estructurada. Además, la Figura 2.6b presenta el proceso de extracción del modelo ER mediante ingeniería inversa de la BDR, el cual se divide en tres principales etapas: *a)* se establece una conexión a la BDR, *b)* se obtiene la información de las entidades atributo y *c)* se procesa la información entidad-relación.

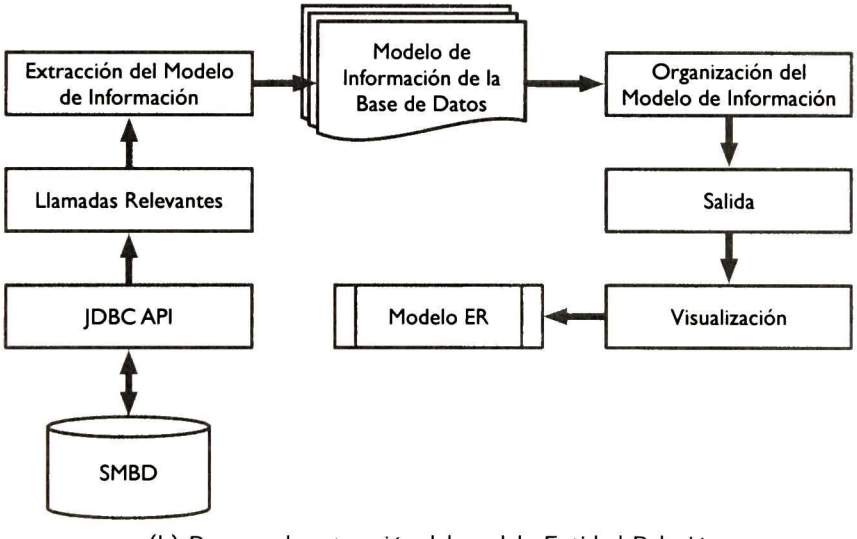
2.3.3 Obtención de TMs a partir de BDRs

Por otro lado, la técnica del modelado del EBD consiste en asignar el esquema a un TM de acuerdo a un conjunto de reglas de traducción predefinidas. No obstante, esta técnica se suele mezclar con la técnica de ingeniería inversa en el proceso de construcción de TMs. En este sentido, Eslami *et al.* [Eslami and Nazami, 2011] presentan un enfoque semiautomático para la construcción de TMs a partir de BDRs. El TM se construye de forma metódica, primero se construyen los tipos de tópicos (*topic types*) y después se añaden sus respectivas instancias. A los tipos de tópico, asociaciones y ocurrencias necesarios para modelar un determinado recurso de información se le denomina ontología de un TM. La ontología básica consiste de cuatro tipos de tópicos:

1. Tipos de tópicos,
2. Tipos de asociaciones,
3. Tipos de roles de asociaciones y
4. Tipos de ocurrencias.



(a) Capa de índice de TMs



(b) Proceso de extracción del modelo Entidad-Relación

Figura 2.6: Enfoque propuesto por Fei *et al.* [Ye et al., 2011]

El modelado de tópicos, asociaciones y ocurrencias se realiza acorde a los tipos de tópicos listados y de acuerdo a su definición en el *Modelo de Datos de Topic Maps* (MDTM). El modelado de los tipos de asociaciones y sus respectivos tipos de roles se realiza con base a la declaración de distintos tipos de relaciones establecidas en la BDR. Por otro lado, la definición de los tipos de ocurrencia se realiza a partir de un archivo de configuración que contiene un listado de los atributos en la BDR a modelarse. Es decir, se debe tener conocimiento a *priori* de los tipos de datos que contiene la BDR y así determinar qué atributos deben modelarse como tipos de ocurrencia. El prototipo propuesto por Eslami *et al.* [Eslami and Nazami, 2011] toma como entrada: *a)* el EBD y el IBD, y *b)* un archivo de configuración con los tipos de ocurrencias en la BDR. Los TMs resultantes se validan de manera sintáctica mediante el editor estándar TM4L[9] [Dicheva and Dichev, 2006].

Un TM busca satisfacer el principio fundamental de organización para la creación y mantenimiento de la información, mediante una red de enlaces semánticos que relacionen diferentes recursos de información. En este sentido, los enfoques reportados evidencian al menos una de las siguientes desventajas:

- convierten la BDR sin considerar el conocimiento implícito en la base de datos,
- requieren de un archivo de configuración donde se especifica cómo la información en la BDR debe traducirse al componente más indicado del TM, y
- carecen de una validación semántica de los TMs resultantes.

2.3.4 Diseño de TMs a partir de recursos de información

En este apartado se presentan algunos enfoques reportados en la literatura para la construcción de TMs. El diseño de TMs requiere de al menos tres fases principales:

1. La primera de ellas se centra en la preparación de los recursos de conocimiento (corpus de documentos, colección de páginas web y documentos XML).

[9]Disponible en: http://compsci.wssu.edu/iis/nsdl/download.html

2. La segunda fase consiste en identificar los tipos de tópicos, es decir, los tópicos que definen las clases de los tópicos (tipos de tópicos) por ejemplo *"Persona"*, *"Compañía"*, asociaciones como *"empleadoPor"*, y ocurrencias como *"Sitio web"*. Después, se deben identificar los tópicos (instancias de los tipos de tópicos) por ejemplo, *"María"*, *"Oracle"*, identificación de asociaciones *"María es empleada de Oracle"* e identificación de ocurrencias por ejemplo, la página web de Oracle o el correo electrónico de María.

3. La fase final se refiere a la evaluación y validación del Topic Map.

Existen distintos trabajos relacionados con el diseño de TMs a partir de recursos de información. La Tabla 2.2 presenta un resumen de los principales enfoques propuestos en la literatura especializada de éstos trabajos. Algunos enfoques han sido propuestos para el diseño, gestión y mantenimiento de TMs. Se observa que los enfoques reportados en la literatura para construir y poblar un TM parten de un recurso de información existente y emplean un conocimiento a *priori* (tal como ontologías de dominio o tesauros). Estos enfoques toman como entrada distintos tipos de datos: documentos estructurados y no estructurados [Reynolds and Kimber, 2000], bases de datos [Eslami and Nazami, 2011], datos semiestructurados [Lin and Qin, 2002]. Existen distintas fuentes de datos que pueden mapearse de manera directa a un TM como metadatos RDF [Gronmo, 2002]. Observamos también que supervisar el proceso de construcción añade un valor considerable al TM resultante, de hecho, la mayoría de los enfoques reportados son una combinación de auto-generación y enriquecimiento manual.

Algunos enfoques [Habert and Folch, 2002, Lavik and Nordeng, 2007, Kasler et al., 2006] proponen el uso de técnicas de aprendizaje y procesamiento en lenguaje natural (PLN) para la obtención de tópicos y asociaciones a partir del texto de documentos. Los métodos de aprendizaje se pueden aplicar con diferentes niveles de automatización: manual, semiautomáticos, automáticos. Algunos trabajos de investigación [Lin and Qin, 2002, Librelotto et al., 2004, Jung-Mn et al., 2007] se enfocan al diseño cooperativo y fusión de TMs.

Por otro lado, también existen algunos enfoques [Laclavík, 2006, Ching-Song Don Wei, 2011, Heru Agus Santoso, 2011] para la creación de Ontologías a partir de BDRs. Estos enfoques extraen la información a partir del BDRs y transforman los datos relacionales a archivos OWL. Además

existen enfoques para la transformación de la BDRs a un marco de descripción de recursos (RDF por sus siglas en inglés) [Bizer and Cyganiak, 2004, Hai-yun and Shu-feng, 2010], que es un lenguaje capaz de ser procesado por una computadora y se utiliza para describir diferentes recursos web. Hai-yun [Hai-yun and Shu-feng, 2010] proponen un enfoque basado en reglas (por el usuario) para la transformación automática del EBD.

Tabla 2.2: Resumen de los enfoques propuestos para la construcción de TMs

ENFOQUE	OBJETIVO PRINCIPAL	TECNICA	REUTILIZACIÓN	RECURSO	HERRAMIENTA	EVALUACIÓN
Gronmo [Gronmo, 2002]	Generación de TMs a partir de declaraciones RDF	- Mapeo de RDF a TMs	No	- Declaración de RDF en BDR	*[a]	*
Reynolds and Kimber [Reynolds and Kimber, 2000]	Creación de TMs a partir de ontologías	- XSLT - TM merging	Ontologías	- Recursos XML	TM4J	- Usuario - Experto
Lin and Qin [Lin and Qin, 2002]	Diseñar un repositorio de conocimientos para integrar y compartir estructuras de conocimiento existentes	- Mejoramiento de TMs con estructuras jerárquicas	No	- Tesauros - Ontologias - Topic Maps	- Hojas de estilo XML para la visualización de TMs	- Trabajo futuro
Librelotto [Librelotto et al., 2004]	TM Builder: Construcción automatica de TMs a partir de documentos XML	- Extracción automática de TMs	No	- Documentos XML	- Constructor de TMs	- Usuario
Roberson y Dicheva [Roberson and Dicheva, 2007]	Extracción semi-automática para crear TMs	- Crawling de sitios web	Ontologías	- Páginas web	- TM4J - TM4L - WebSphinx	- Usuario
Diacheva y Dichev [Dicheva and Dichev, 2006]	TML4: Entorno para la creación y navegación de TMs educativos	- Clasificación (estructura ontológica para indexar el contenido del repositorio)	Ontologías del dominio	- Repositorios de aprendizaje	- TM4L	- Usuario experto
Kasler *et al.* [Kasler et al., 2006]	Generación semi-automática de TMs a partir de documentos de texto	- Aprendizaje automático - Mapeo de metadatos a TMs	Ontologías del dominio	- Texto - Metadatos estructurados	- Framework Tapestry	*

[a] Información no proporcionada por los autores.

3

Metodología

El objetivo principal de este capítulo es presentar una metodología basada en reglas para la obtención automática de Topic Maps (TMs) a partir de Bases de Datos Relacionales (BDRs).

El *Modelo de Datos de Topic Maps* (MDTM) define de manera formal los principales elementos semánticos de un TM, estos son los *tópicos*, *asociaciones* y *ocurrencias*. A esta triada de elementos se le conoce como el TAO de los TMs. Con base a este conjunto de elementos TAOs, la ontología básica propuesta consiste de cuatro tipos de tópicos:

- tipos de tópicos (*topic types*),
- tipos de asociaciones (*association types*),
- tipos de roles en asociaciones (*association role types*) y
- tipos de ocurrencias (*occurrence types*).

El enfoque propuesto considera como entrada: el *Esquema de la Base de Datos* (EBD), el *Estado de la Base de Datos* (IBD), las *Restricciones de Integridad* (RIs) y un conjunto de tipos de asociaciones y roles. La Figura 3.1 muestra un esquema general del proceso de transformación

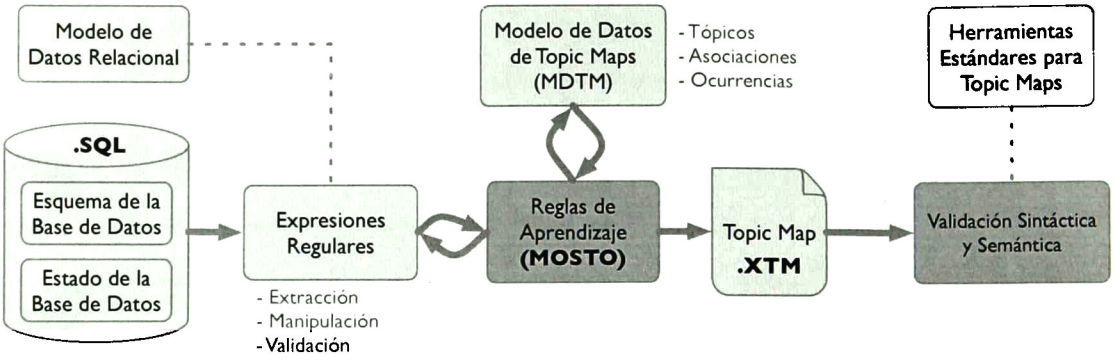


Figura 3.1: Esquema general para la obtención de TMs a partir de datos relacionales

para la obtención del TM a partir de una BDR. El proceso de extracción, manipulación y validación de datos se realiza mediante Expresiones Regulares (ERs) acorde al MDR. Los datos extraídos son manipulados por un conjunto de reglas de aprendizaje (sistema MOSTO), las cuales se encargan de validar y representar los datos en los componentes más idóneos del estándar TM. Así, todo el proceso de transformación es guiado por las reglas de aprendizaje de acuerdo a la definición de los elementos semánticos en el MDTM. Finalmente, el TM obtenido es validado de distintas formas por diferentes herramientas estándares reportadas en la literatura especializada (ver Capítulo 5).

Recordemos que el **esquema de la base de datos relacional** S es un conjunto de *esquemas de relación* $S = \{R_1, R_2, \ldots, R_m\}$ y un conjunto de Restricciones de Integridad RI. Mientras que el **estado de la base de datos relacional**[1] E obtenido a partir de S, es un conjunto de *estados de relación* $E = \{r_1, r_2, \ldots, r_m\}$, tal que cada r_i es un estado de R_i y tal que cada estado de relación r_i satisface a las restricciones de integridad especificadas en RI. Donde un **esquema de relación** R se compone de una lista de atributos, $A_1, A_2, ..., A_n$. Cada atributo A_i es el nombre de un rol desempeñado por algún dominio denotado por $dom(A_i)$. Por otro lado, un **estado de relación**[2] r es un conjunto de *n*-tuplas $r = \{t_1, t_2, \ldots, t_p\}$. Cada tupla t es una lista ordenada de n valores $t = \langle v_1, v_2, \ldots, v_n \rangle$, donde cada valor v_i es un elemento de $dom(A_i)$. El *i*-ésimo valor en la tupla t, que corresponde al atributo A_i, es referenciado como $t[A_i]$.

[1] El EBD es también denominada base de datos de instancias.
[2] También se le llama relación de instancias.

3.1 Transformación de datos relacionales

El proceso de transformación de los datos relacionales es guiado por un conjunto de reglas de aprendizaje. Las reglas de aprendizaje permiten transformar datos relacionales a los elementos que integran el MDTM. El MDTM define de manera formal los elementos semánticos de un TM, donde los principales elementos son los tópicos, asociaciones y ocurrencias. Se proponen en total seis reglas de aprendizaje principales, las primeras tres permiten transformar los tipos de tópicos, asociaciones y tipos de ocurrencias. Otras 3 reglas permiten transformar las instancias de cada tipo de tópico, asociaciones y ocurrencias. Además, cada regla posee distintos tipos de restricciones y propiedades que hacen posible la transformación exitosa de los datos relacionales. Estos tipos de restricciones y propiedades son detallas y ejemplificadas en las secciones siguientes.

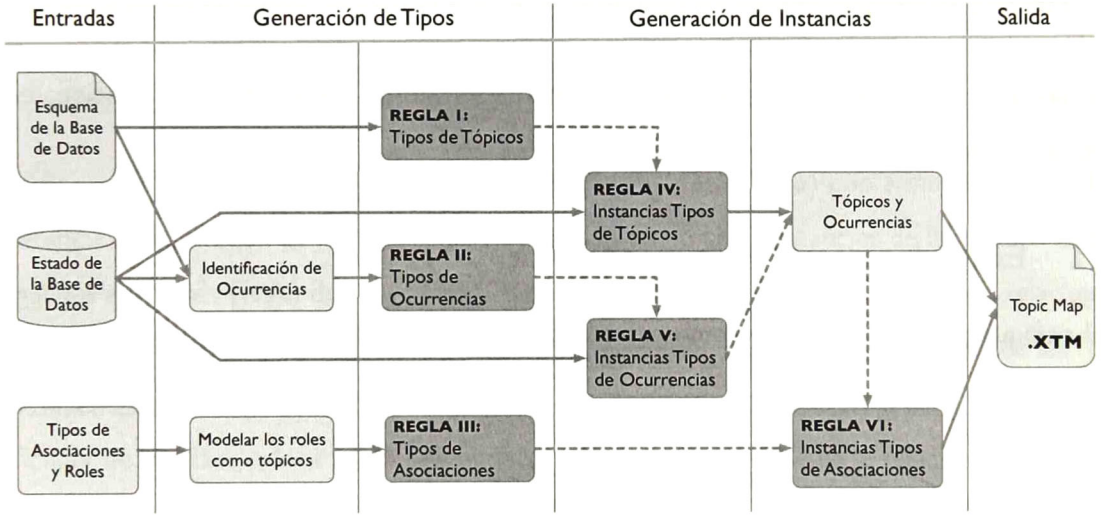


Figura 3.2: Interacción de las reglas de aprendizaje en el proceso de transformación

La Figura 3.2 presenta las seis reglas de aprendizaje propuestas y la manera en como éstas interactúan durante el proceso de transformación de los datos relacionales. Se observa que el proceso de transformación se encuentra dividido en cuatro partes principales: a) las entradas al proceso de transformación, b) el proceso de generación de los tipos de tópicos, c) el proceso de generación de las instancias de los tipos de tópicos y d) la salida del proceso de transformación (TM resultante).

Se han definido una serie de funciones para la explicación de la metodología propuesta. La función A = attr(R) obtiene la lista de atributos del esquema de relación R, la función D = dom(A_i) obtiene un valor válido para el atributo $A_i \in A$. Las funciones P = pkey(R) y F = fkey(R) obtienen las llaves primarias y llaves foráneas respectivamente. Finalmente, la función O = occ(R) representa a los atributos que deben modelarse como tipos de ocurrencias (*occurrence types*) para el esquema de relación R (ver Sección 3.1.2). En este contexto, la Figura 3.3 muestra los conjuntos previamente definidos, donde se cumple que $P \subset A$, $F \subset A$ y $O \subset A$.

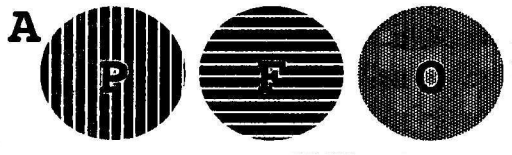


Figura 3.3: Visualización de los conjuntos A, P, F y O para un esquema de relación R

En las secciones siguientes se describe el proceso de transformación de los datos relacionales para la obtención de un TM.

3.1.1 Entradas al enfoque propuesto

El enfoque propuesto requiere como entrada tres elementos principales:

- El *Esquema de la Base de Datos* (EBD) y su respectivo conjunto de *Restricciones de Integridad* (RIs).
- La instancia o *Estado de la Base de Datos* (IBD)
- Un conjunto de tipos de asociaciones y roles para establecer las relaciones semánticas entre los tópicos.

Esquema de la Base de Datos Relacional

Para la demostración del enfoque propuesto se utiliza la BDR SALES descrita a continuación. La Figura 3.4 muestra el EBD y las RIs de SALES = {SHIPPERS, ORDERS, PRODUCTS,

ORDER_DETAILS}. Los atributos subrayados representan llaves primarias. En los siguientes apartados se utiliza este EBD SALES para ejemplificación y demostración de la metodología propuesta.

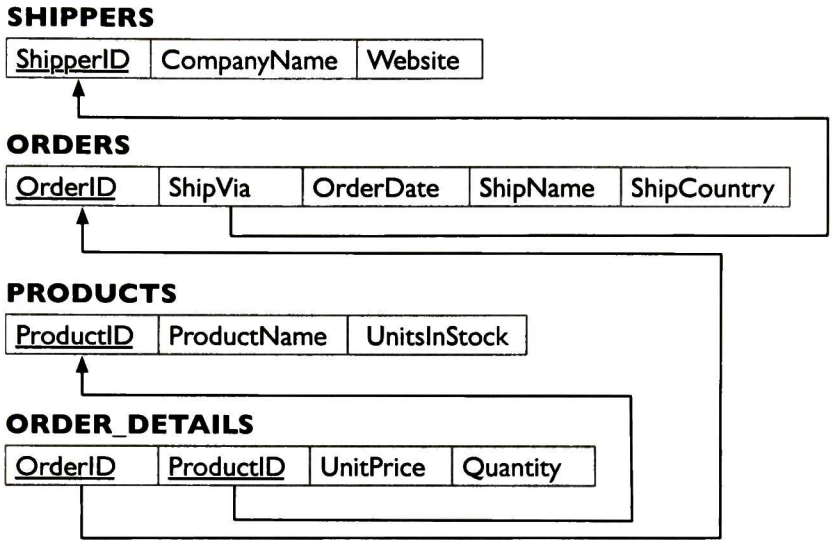


Figura 3.4: EBD y RIs de referencia de la BDR SALES

El EBD de la BDR SALES contiene dos tipos de RIs de referencia: a) *uno a muchos* (1:N) donde una instancia de la relación SHIPPERS está relacionada con muchas instancias de la relación ORDERS y b) *muchos a muchos* (N:N) donde muchas instancias de la relación PRODUCTS están relacionadas con muchas instancias de la relación ORDERS. Otro factor importante que se considera de las RIs de referencia es el cambio de nombre en los atributos entre una relación y otra. Por ejemplo, el atributo identificador de la relación SHIPPERS se denomina ShipperID mientras que la relación ORDERS hace referencia al mismo atributo renombrado como ShipVia. Además de las RIs de referencia, el IBD SALES incluye RIs de domino. Por ejemplo, el atributo UnitsInStock no debe tener asociado un número *flotante* o un número entero negativo.

Estado de la Base de Datos Relacional

Además del EBD, el enfoque propuesto requiere como entrada el IBD. La Figura 3.5 muestra el IBD que cumple con el EBD SALES (ver Figura 3.4). El IBD SALES incluye distintos tipos de

datos con su respectivo dominio. Por ejemplo, cadenas de caracteres, valores nulos, números enteros y flotantes, fechas y enlaces a páginas web. En los siguientes apartados se utiliza este IBD SALES para ejemplificación y demostración de la metodología propuesta.

ORDERS

OrderID	ShipVia	OrderDate	ShipName	ShipCountry
100	10	8/11/1997	Ernst Handel	NULL
101	11	9/22/1997	Die Wandernde Kuh	Germany
102	10	8/21/1997	Emparedados y helados	Mexico

ORDER DETAILS

OrderID	ProductID	UnitPrice	Quantity
100	51	45.90	12
101	50	99.34	55
102	51	45.90	30

PRDUCTS

ProductID	ProductName	UnitsInStock
50	Chocolate	200
51	Filo Mix	23

SHIPPERS

ShipperID	Company Name	Website
10	Speed Express	http://www.speedexpress.com
11	United Package	http://www.unitedpackage.com

Figura 3.5: IBD para el EBD SALES

Tipos de asociaciones propuestas

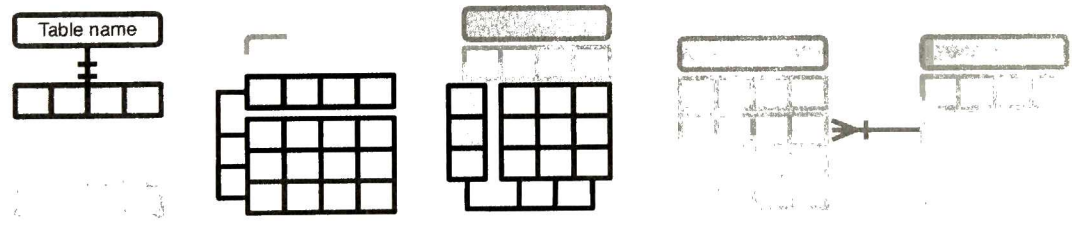
Las asociaciones en el estándar TM representan relaciones entre tópicos y éstas pueden ser etiquetadas o nombradas. Las relaciones en una BDR se representan mediante llaves foráneas mientras que en el MDTM se representan como un tipo de asociación (*association type*). Para cada tipo de asociación se definieron dos roles, donde un rol define el papel que desempeña un determinado tópico en el tipo de relación. Los tópicos (elemento principal) poseen tres principales características: un nombre, ocurrencias y desempeña un determinado rol en una asociación. En la Figura 3.6 se muestran los elementos de un tipo de asociación:

Figura 3.6: Nombre de los elementos que conforman un tipo de asociación

La Tabla 3.1 presenta los tipos de asociaciones propuestas con sus respectivos roles, mientras que la Figura 3.7 los presenta de manera gráfica.

Tabla 3.1: Definición de los tipos de asociaciones y roles

TIPO DE ASOCIACIÓN	ROL PRIMARIO	ROL SECUNDARIO
is-attribute-of	Superclass	Attribute
is-instance-of	Class	Instance
has-a-relation-to	Primary-key	Relation-to
has-a-reference-to	Reference	Foreign-Key



(a) *is-attribute-of* (b) *is-instance-of* (c) *has-a-relation-to* (d) *has-a-reference-to*

Figura 3.7: Representación gráfica de los tipos asociaciones propuestos

A continuación se presenta una descripción de los cuatro tipos de asociación necesarios y requeridos para establecer las relaciones semánticas entre los tópicos:

Es-atributo-de (is-attribute-of) indica una asociación entre el nombre de un esquema de relación R y su respectiva lista de atributos denotado por A, donde el rol que desempeña R se denomina *superclase (superclass)* y el rol de cada atributo $A_i \in A$ se denomina *atributo (attribute)*. La Figura 3.8 muestra de forma gráfica la transformación de los datos relacionales a su representación semántica mediante el tipo de asociación *es-atributo-de* con sus respectivos roles.

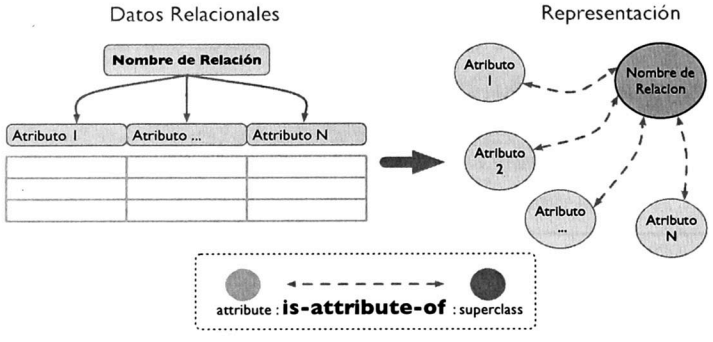


Figura 3.8: Representación gráfica del tipo de asociación *es-atributo-de*

Es-instancia-de (is-instance-of) indica una asociación entre un atributo $A_i \in A$ y sus respectivas instancias o estados $t[A_i]$, donde el rol que desempeña cada A_i se denomina *clase (class)* y el rol de cada valor $t[A_i]$ se denomina *instancia (instance)*. La Figura 3.9 muestra de forma gráfica la transformación de los datos relacionales a su representación semántica mediante el tipo de asociación *es-instancia-de* con sus respectivos roles.

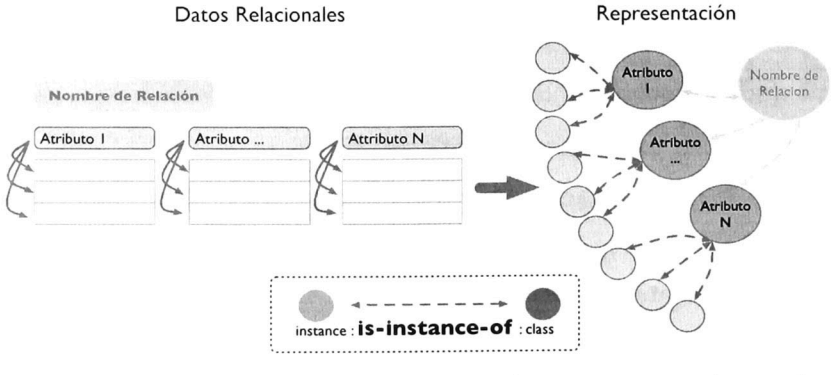


Figura 3.9: Representación gráfica del tipo de asociación *es-instancia-de*

Tiene-relación-con (has-a-relation-to) indica una asociación entre la instancia de un atributo considerado como llave primaria $t[A_1]$ y sus instancias asociadas $t = \langle v_2, v_3, \ldots, v_n \rangle$, donde el rol que desempeña $t[A_1]$ se denomina *llave-primaria (primary-key)* y el rol de cada instancia $t[A_i] \in t$ se denomina *relacionado-a (relation-to)*. La Figura 3.10 muestra de forma gráfica la transformación de los datos relacionales a su representación semántica mediante el tipo de asociación *tiene-relación-con* con sus respectivos roles.

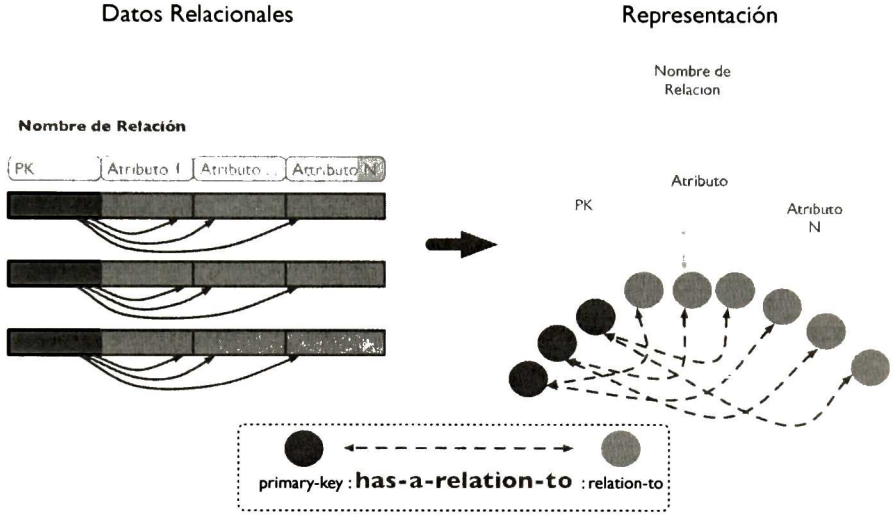


Figura 3.10: Representación gráfica del tipo de asociación *tiene-relación-con*

Tiene-una-referencia-a (has-a-reference-to) indica una asociación de las instancias de atributos considerados como llaves primarias entre dos o más relaciones, esto de acuerdo a las RIs de referencia de la BDR en cuestión, donde la instancia del atributo llave primaria desempeña el rol de *referencia (reference)* y el rol de cada instancia del atributo al que se hace referencia se denomina *llave-foránea (foreign-key)*. La Figura 3.11 muestra de forma gráfica la transformación de los datos relacionales a su representación semántica mediante el tipo de asociación *tiene-relación-con* con sus respectivos roles.

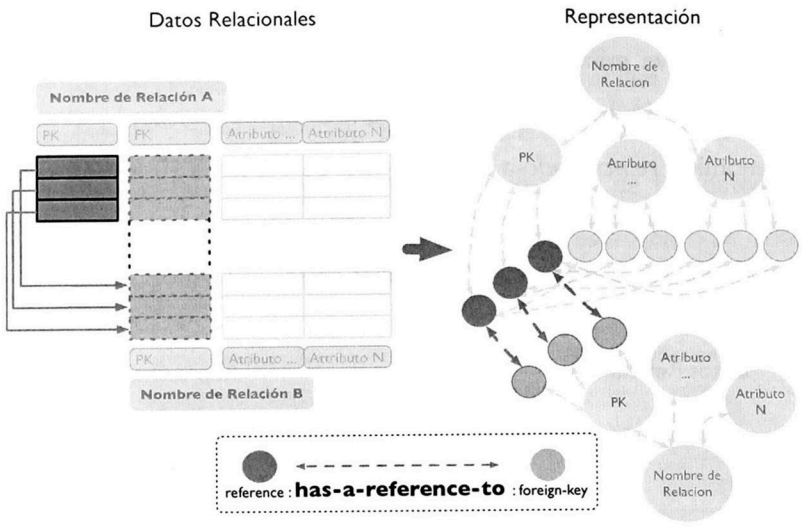


Figura 3.11: Representación gráfica del tipo de asociación *tiene-referencia-con*

3.1.2 Generación de tipos de tópicos

La generación de los tipos de tópicos consta de tres principales procesos, por lo cual se proponen las reglas siguientes:

Regla I: Generación de tipos de tópicos

Debe modelarse como un tipo de tópico (*topic type*):

- Cada nombre de relación R en S.

  Por ejemplo, todos los elementos en el EBD de SALES = {SHIPPERS, ORDERS, PRODUCTS, ORDER_DETAILS} deben modelarse como tipos de tópicos. A continuación se presenta la sintaxis de la creación del tipo de tópico *PRODUCTS* a partir de la relación PRODUCTS:

```
<topic id="products">          Identificador del tópico
  <name>
     <value>Products</value>   Valor del tópico
  </name>
</topic>
```

- Cada elemento en el conjunto T para cada R que satisfaga:

$$T = \begin{cases} A - (P \cup F \cup O) & \text{if } F = P \\ A - (F \cup O) & \text{else} \end{cases} \tag{3.1}$$

Una lista de atributos A puede contener hasta cuatro tipos de atributos: a) atributos que representan llaves primarias, b) atributos que representan llaves foráneas c) atributos que deben modelarse como tipos de ocurrencias y d) atributos "normales" que no pertenecen a los otros tipos de atributos. Los atributos en el conjunto T que satisfacen la Ecuación 3.1 deben modelarse como tipos de tópicos. La Figura 3.12 muestra las posibles variantes de los los conjuntos P y F, si se cumple que:

- la intersección entre P y F es el conjunto vacío o
- P y F comparten algunos elementos o
- F es un subconjunto propio de P

entonces $T = A - (F \cup O)$.

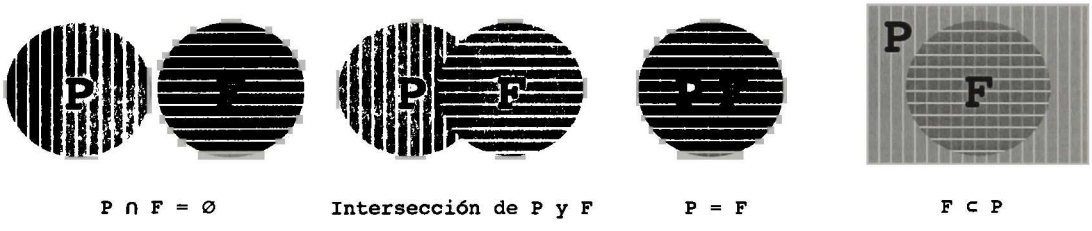


Figura 3.12: Variantes de los conjuntos P y F en una lista de atributos A

Por ejemplo, para la relación ORDERS donde A = {OrderID, ShipVia, OrderDate, ShipName, ShipCountry} se tiene que P = {OrderID}, F = {ShipVia} y O = {}. Para la relación ORDERS se cumple que la intersección de los conjuntos P y F es el conjunto vacío. Por lo tanto de acuerdo a la Ecuación 3.1 se satisface $T = A - (F \cup O)$ y los elementos T = {OrderID, OrderDate, ShipName, ShipCountry} deben modelarse como tipos de tópicos.

A continuación se presenta la sintaxis para la creación del tipo tópico *ShipName* a partir del atributo ShipName de la relación ORDERS:

```
<topic id="shipname-orders">   Identificador del tópico
  <name>
     <value>ShipName</value>  Valor del tópico
  </name>
</topic>
```

Además, existe una restricción cuando F = P La Ecuación 3.1 indica que deben eliminarse los conjuntos F y P de A. La restricción indica que de acuerdo al Modelo de Datos Relacional, se trata de una Restricción Integridad de referencia de *muchos-a-muchos*. Si esta restricción de referencia se cumple, entonces debe crearse un nuevo tipo de tópico *maestro*.

Por ejemplo, para la relación ORDER_DETAILS donde A = {OrderID, ProductID, UnitPrice, Quantity} se tiene que P = {OrderID, ProductID}, F = {OrderID, ProductID} y O = {}. Para la relación ORDER_DETAILS se cumple que los conjuntos F y P son iguales. Por lo tanto de acuerdo a la Ecuación 3.1 se satisface T = A − (P ∪ F ∪ O) y los elementos T = {UnitPrice, Quantity} deben modelarse como tipos de tópicos, y debe crearse un nuevo tipo de tópico *maestro* para la relación ORDER_DETAILS:

```
1 <topic id="mpk-orderdetails">          Identificador
2   <name>
3      <value>MPK Order Details</value> Valor del tópico
4   </name>
5 </topic>
```

Regla II: Generación de tipos de asociaciones

Debe modelarse:

- Cada rol primario y secundario de los tipos de asociaciones propuestas debe modelarse como un tópico (ver Tabla 3.1). A continuación se muestra la sintaxis del rol primario *Class* para el tipo de asociación *es-instancia-de* (is-instance-of):

```
1 <topic id="class-primaryrole">  Identificador del tópico
2   <name>
3     <value>Class</value>        Nombre del rol
4   </name>
5 </topic>
```

- Cada tipo de asociación propuesto debe modelarse como un tópico de acuerdo a la plantilla presentada en la Figura 3.13.

```
1 <topic id="is-ttribute-of">                                  Identificador del tópico
2   <name>
3     <value>is attribute of</value>                           Nombre del tipo de asociación
4   </name>
5   <name>
6     <scope><topicRef href="#superclass"/></scope>            Identificador del rol secundario
7     <value>is main of</value>                                Etiqueta de jerarquía primaria
8   </name>
9   <name>
10    <scope><topicRef href="#attribute"/></scope>             Identificador del rol primario
11    <value>is instance of</value>                            Etiqueta de jerarquía secundaria
12  </name>
13 </topic>
```

Figura 3.13: Plantilla para el modelado de los tipos de asociación (ejemplo: *es-atributo-de*)

Regla III: Generación de tipos de ocurrencias

Las ocurrencias son recursos de información relevante que ayudan a definir un determinado tópico. Existen distintos tipos de ocurrencias que permiten diferenciar entre las variantes de relaciones a los recursos de información. Algunos ejemplos de tipos de ocurrencias son el correo electrónico, una URI (identificador uniforme de recurso), una fórmula, una fecha, una definición conceptual, etc. En el enfoque propuesto los tipos de ocurrencias se identifican de manera automática a partir de los atributos de una BDR.

El Algoritmo 1 propuesto para la identificación automática de tipos de ocurrencias requiere como entrada tres parámetros:

1. El esquema de relación R donde se define su lista de atributos $A_1, A_2, \ldots, A_n$.

2. El estado de relación $r = \{t_1, t_2, \ldots, t_m\}$ donde se define las instancias de R.

3. El parámetro k, el cual indica el número de tuplas t a explorar.

Algoritmo 1 Identificación de tipos de ocurrencias en BDRs

```
1: Input: Un esquema de relación R, el estado de relación r y el número k de tuplas t a explorar.
2: Output: El conjunto tipos de ocurrencias O en R.

3: For i = 1 to k tuples in r:
4:     For each A_j in R:
5:         instancia := t_i[A_j]
6:         If instancia = NULL then
7:             isNull := True
8:         Else If isOcurrence(instance) then
9:             inject(O, A_j)
10:            isOcurrence := True
11:    If isOcurrence and not isNull then
12:        break
```

El algoritmo produce como salida el conjunto O, el cual contiene los atributos del esquema de relación R, que son considerados como tipos de ocurrencias.

La regla en cuestión señala que cada elemento en el conjunto O para una relación R debe modelarse como un tipo de ocurrencia.

Por ejemplo, para el esquema de relación SHIPPERS, donde A = {ShipperID, ShipperName, Website}, el algoritmo identificó como tipos de ocurrencias a los elementos en O = {Website}. A continuación se presenta la sintaxis para la creación del tipo de ocurrencia *Website* a partir del atributo Website de la relación SHIPPERS:

```
<topic id="website-shippers">  Identificador del tópico
  <name>
    <value>Website</value>  Valor del tópico
  </name>
</topic>
```

3.1.3 Generación de instancias de tipos de tópicos

Después de la generación de los tipos de tópicos, ahora corresponde a la generación de las instancias de los tipos de tópicos. Esta etapa consta de tres principales procesos, por lo cual se proponen las reglas siguientes.

Regla IV: Generación de instancias de los tipos de tópicos

- Dado un estado de relación $r = \{t_i, t_2, \ldots, t_p\}$, para cada tupla $t = \langle v_2, v_3, \ldots, v_n \rangle$ el elemento $t[A_i]$ debe modelarse como una instancia del atributo A_i sí y sólo sí se cumple que:

$$t[A_i] \in dom(A_i) \wedge A_i \in T \tag{3.2}$$

donde T representa al conjunto de atributos modelados como tipos de tópicos, obtenido con la REGLA I. Es decir, si el elemento $t[A_i]$ representa un valor válido (no nulo y que pertenece a su dominio) y que además el atributo A_i en cuestión se haya modelado previamente como un tipo de tópico.

Por ejemplo, las instancias del atributo ShipCountry = {NULL, Germany, Mexico} de la relación ORDERS, la instancia t_1[ShipCountry] = NULL no satisface la Ecuación 3.2 debido a que NULL $\notin$ dom(ShipCountry).

A continuación se presenta la sintaxis para la creación del tópico *Mexico* a partir de la instancia Mexico del atributo ShipCountry:

```
1 <topic id="idmexico">                                  Identificador del tópico
2   <instanceOf>
3      <topicRef href="#shipcountry-orders</value> Referencia al tipo de tópico
4   </instanceOf>
5   <name>
6      <value>Mexico</value>                             Valor del tópico/instancia
7   </name>
8 </topic>
```

Regla V: Generación de instancias de los tipos de asociaciones

Deben modelarse todas las instancias de los 4 tipos de asociaciones propuestas siguiendo las siguientes restricciones.

1. *Es-atributo-de*: Dado los conjuntos A y F de un esquema de relación R, debe modelarse una instancia del tipo de asociación *is-attribute-of* entre el nombre de la relación R y cada atributo A_i resultante de la condición siguiente:

$$\{A_i \mid A_i \in A \wedge A_i \notin F\}. \tag{3.3}$$

   Por ejemplo, para el esquema de relación ORDERS, donde A = {OrderID, ShipVia, OrderDate, ShipName, ShipCountry} y F = {ShipVia}, los elementos resultantes al aplicar la Ecuación 3.3 son {OrderID, OrderDate, ShipName, ShipCountry}.

   A continuación se presenta la sintaxis para la creación de la instancia del tipo de asociación *is-attribute-of* entre la relación ORDERS y el atributo ShipName:

```
<association>
   <type>
      <topicRef href="#is-attribute-of"/>          Identificador del tipo de asociación
   </type>
   <role>
      <type><topicRef href="#superclass"/></type>  Identificador del rol secundario
      <topicRef href="#orders"/>                   Identificador de la relación
   </role>
   <role>
      <type><topicRef href="#attribute"/></type>   Identificador del rol primario
      <topicRef href="#shipname-orders"/>          Identificador del atributo
   </role>
</association>
```

2. *Es-instancia-de*: Dado un esquema de relación R y su respectivo estado de relación r, debe modelarse una instancia del tipo de asociación *is-instance-of* entre un atributo A_i y cada valor $t[A_i]$, donde el atributo A_i debe satisfacer la Ecuación 3.3 y se debe cumplir que el valor $t[A_i] \in \text{dom}(A_i)$.

Por ejemplo, a continuación se presenta la sintaxis para la creación de la instancia del tipo de asociación *is-instance-of* entre el atributo ProductName y su instancia "Chocolate":

```
1 <association>
2   <type>
3      <topicRef href="#is-instance-of"/>          Identificador del tipo de asociación
4   </type>
5   <role>
6      <type><topicRef href="#class"/></type>      Identificador del rol secundario
7      <topicRef href="#productname"/>             Identificador del atributo
8   </role>
9   <role>
10     <type><topicRef href="#instance"/></type>   Identificador del rol primario
11     <topicRef href="#idchocolateproductname"/>  Identificador de la instancia
12  </role>
13 </association>
```

3. *Tiene-relación-con*: Dado el estado de relación r de un esquema de relación R, debe modelarse una instancia del tipo de asociación *has-a-relation-to* entre la instancia de un atributo considerado como llave primaria $t[A_1]$ y cada una de sus instancias $t = \langle v_2, v_3, \ldots, v_n \rangle$, donde el atributo A_i debe satisfacer la Ecuación 3.3 y se debe cumplir que el valor $t[A_i] \in dom(A_i)$.

   Por ejemplo, para la relación PRODUCTS cuando $t_1 = \langle 50, \text{Chocolate}, 200 \rangle$ donde $t[A_1] = 50$ y $t = \langle \text{Chocolate}, 200 \rangle$. A continuación se presenta la sintaxis para la creación de la instancia ... tipo de asociación *has-a-relation-to* entre el valor considerado como llave primaria "50" y su valor ...ociado "Chocolate":

```
1 <associatio...
2   <type>
3      <topicRef hr...  "#has-a-relation-to"/>      Identificador del tipo de asociación
4   </type>
5   <role>
6      <type><topicRef href...  ...rimary-key"/></type>  Identificador del rol secundario
7      <topicRef href="#id50p...  ...ctid"/>          Identificador de la llave primaria
8   </role>
9   <role>
10     <type><topicRef href="#relation-to"/></type>  Identificador del rol primario
11     <topicRef href="#idchocolateproductname"/>    Identificador de la instancia
12  </role>
13 </association>
```

4. *Tiene-referencia-a*: Dado el estado de relación r de un esquema de relación R, debe modelarse una instancia del tipo de asociación *has-a-reference-to* entre la instancia de un atributo considerado como llave primaria $t[A_1]$ y cada una de las instancias de los atributos considerados como llave foránea $t[A_i]$, donde $t[A_1] \neq t[A_i]$.

   A continuación se presenta la sintaxis para la creación de la instancia del tipo de asociación *has-a-reference-to*, entre el valor considerado como llave primaria t[OrderID] = 100 y el valor considerado como llave foránea t[ShipVia] = 10, de la relación ORDERS:

```
1  <association>
2    <type>
3       <topicRef href="#has-a-reference-to"/>          Identificador del tipo de asociación
4    </type>
5    <role>
6       <type><topicRef href="#reference"/></type>     Identificador del rol secundario
7       <topicRef href="#id100orderid"/>               Identificador de la llave primaria
8    </role>
9    <role>
10      <type><topicRef href="#foreign-key"/></type>   Identificador del rol primario
11      <topicRef href="#id10shipperid"/>              Identificador de la instancia
12   </role>
13 </association>
```

Regla VI: Generación de instancias de los tipos de ocurrencias

- Dado un estado de relación $r = \{t_i, t_2, \ldots, t_m\}$. Para cada tupla $t = \langle v_1, v_2, \ldots, v_n \rangle$, el elemento $t[A_i]$ debe modelarse como una instancia del atributo A_i si y solo si se cumple que:

$$t[A_i] \in dom(A_i) \wedge A_i \in O \tag{3.4}$$

donde O representa al conjunto de atributos modelados como tipos de ocurrencias, obtenido con la REGLA III. La instancia $t[A_i]$ se añade a la llave primaria de t, utilizando el elemento <ocurrence> del MDTM.

Para la relación SHIPPERS donde A = {ShipperID, CompanyName, Website} y se tiene que el atributo Website es un tipo de ocurrencia. Por ejemplo, para la tupla t_1 =<10, Speed

Express, http://www.speedexpress.com> la llave primaria es "10" y la instancia del tipo de asociación es "http://www.speedexpress.com". A continuación se presenta la sintaxis para la creación/modificación del ejemplo previo:

```
1 <topic id="idhipperid10">
2   <instanceOf>
3      <topicRef href="#shipperid-shippers"/>                Identificador del tipo de tópico
4   </instanceOf>
5   <name>
6      <value>10</value>                                     Valor del tópico (llave primaria)
7   </name>
8   <occurrence>
9      <type><topicRef href="#website-shippers"/></type>     Identificador del tipo de ocurrencia
10     <resourceData>http://speedexpress.com</resourceData>  Valor de la ocurrencia
11  </occurrence>
12 </topic>
```

4

Implementación

En este capítulo se presenta la implementación de la metodología propuesta denominada MOSTO y la integración del mismo a una arquitectura por capas para la gestión de Topic Maps (TMs).

La metodología propuesta fue implementada e integrada a una arquitectura por capas. La implementación busca satisfacer dos objetivos principales: a) servir como prueba de concepto para la validación de los TMs obtenidos y b) servir como un prototipo destinado a los usuarios finales y terceras partes (comunidad científica).

A continuación se describe la implementación de:

- la metodología propuesta para la obtención automática de TMs y
- la aplicación web como prueba de concepto de la metodología y gestión de los TMs resultantes.

4.1 Implementación del Enfoque Propuesto

La implementación del enfoque propuesto MOSTO (modelo semántico para el diseño de Topic Maps), la cual fue realizada en el lenguaje de programación JAVA, debido a que las herramientas de validación para los TMs se encuentran implementadas en este lenguaje.

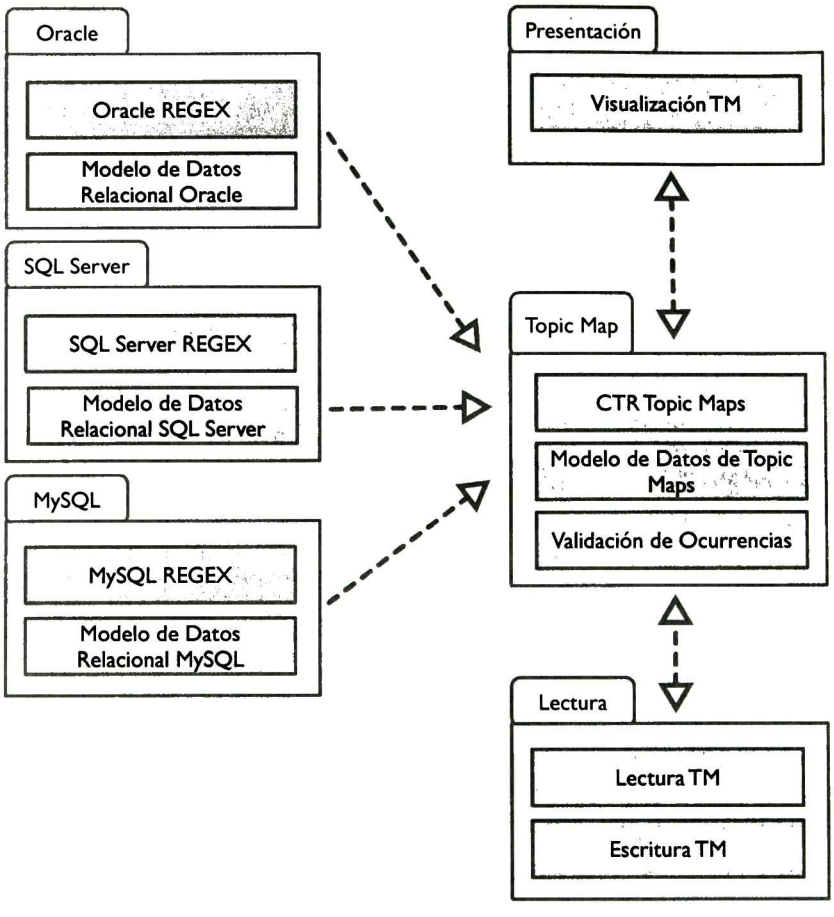


Figura 4.1: Diagrama de paquetes para la obtención de un TM

La implementación sigue los lineamientos del paradigma orientado a objetos. El paradigma utiliza objetos como elementos fundamentales en la implementación de la metodología propuesta. Un objeto es una abstracción de algún hecho del mundo real que tiene atributos que representan sus características y métodos que representan su comportamiento. Estas características y métodos comunes a los objetos se agrupan en clases. Finalmente una clase es una plantilla para crear objetos, por eso los objetos son instancias de clases.

La Figura 4.1 ilustra el diagrama de paquetes donde se visualiza la organización de paquetes y elementos que integran el módulo denominado MOSTO:

Topic Map es el paquete principal, en el se encuentran definidos todos los lineamientos (reglas

propuestas) para la generación de TMs. Además se encarga de la generación de tópicos, asociaciones y ocurrencias, de acuerdo al *Modelo de Datos de Topic Maps* (MDTM). Se relaciona directamente con los paquetes auxiliares:

SQLServer se encarga de extraer, validar y manipular los datos provenientes del *Esquema de la Base de Datos* (EBD). El proceso se realiza mediante *Expresiones Regulares* (ERs) de acuerdo a la especificación del *Modelo de Datos Relacional* (MDR) Microsoft SQL Server®.

MySQL se encarga de extraer, validar y manipular los datos provenientes del EBD . El proceso se realiza mediante ERs de acuerdo a la especificación del MDR MySQL.

Oracle se encarga de extraer, validar y manipular los datos provenientes del EBD. El proceso se realiza mediante ERs de acuerdo a la especificación del MDR Oracle®.

El controlador del paquete se encarga de gestionar los flujos de entrada y salida. Por un lado, la lectura y escritura, y por otro lado la visualización de los TMs obtenidos.

Por otro lado, la Figura 4.2 presenta el diagrama de secuencia para modelar la interacción entre los objetos en el módulo MOSTO. A continuación se describe la interacción entre los elementos del diagrama:

- Primero, crear un objeto *Topic Map* y asociarle un *Sistema Manajador de Base de Datos* (SMBD).
- Crear de manera iterativa objetos *Tipo de Tópico*, esto a partir del *Esquema de la Base de Datos* (EBD). Finalmente agregarlos al *Topic Map*.
- Crear de manera iterativa los objetos que representan una instancia de cada *Tipo de Tópico*, a partir del EBD y de las *Restricciones de Integridad* (RIs). Finalmente agregarlos al *Topic Map*.
- Crear objetos *Tipo de Asociación*, a partir de los cuatro tipos de asociación propuestos (*is-attribute-of*, *is-instance-of*, *has-a-relation-to* y *has-a-reference-to*). Por último agregarlos al *Topic Map*.

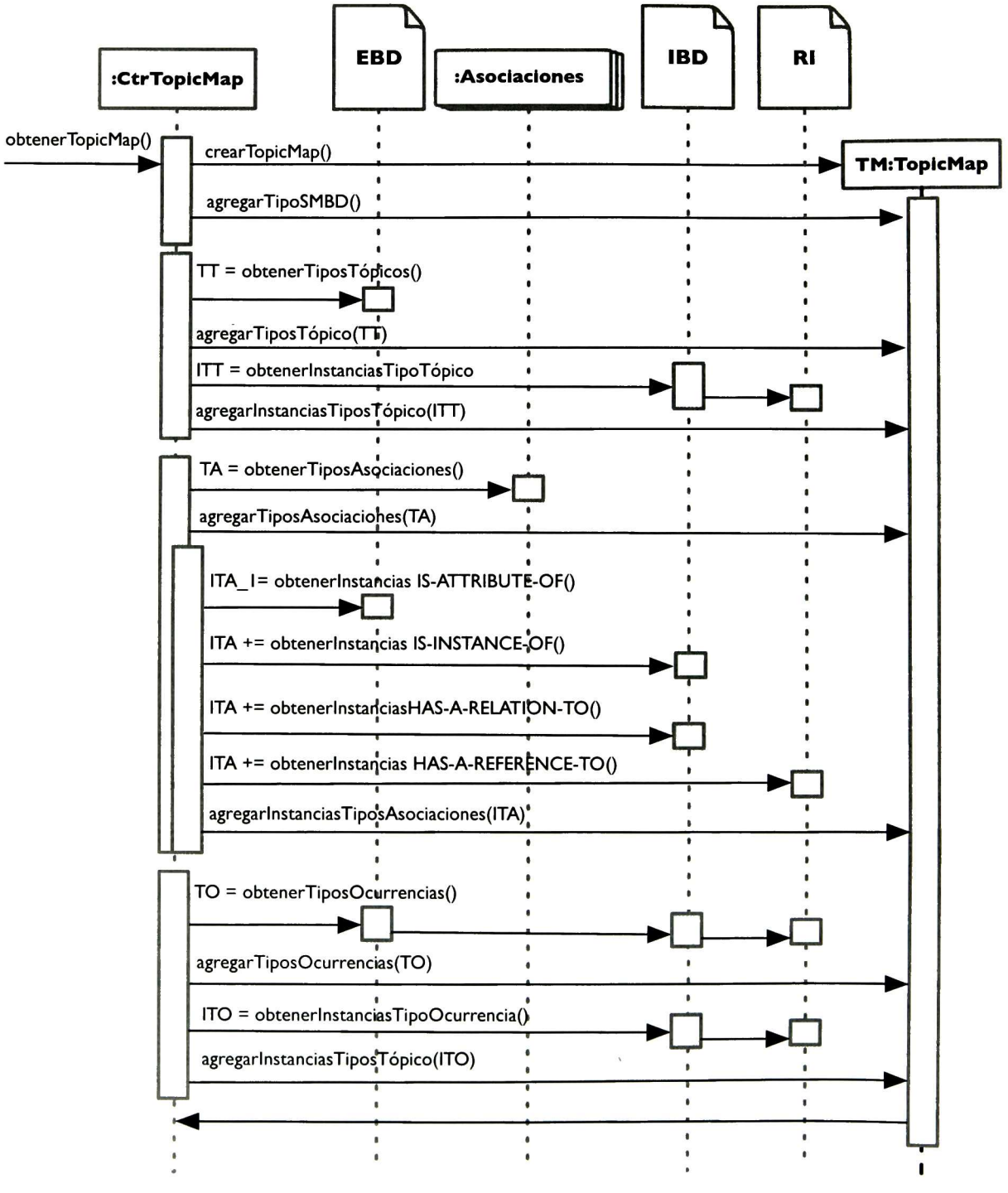


Figura 4.2: Diagrama de secuencia para la obtención de un TM

- Crear de manera iterativa los objetos que representan una instancia de cada tipo de asociación, a partir del EBD, IBD y de las RIs, esto dependiendo del tipo de asociación que se esté modelando. Finalmente, agregar las instancias al *Topic Map*.

- Crear de manera iterativa objetos *Tipo de Ocurrencia*. Determinar qué atributos deben modelarse requiere de la validación de sus instancias en el IBD y también de las RIs. Finalmente agregar los objetos creados al *Topic Map*.

- Crear de manera iterativa los objetos representan una instancia de cada *Tipo de Ocurrencia*, a partir del EBD, IBD y de las RIs. Finalmente agregarlos al *Topic Map*.

- Por último, regresar el *Topic Map* resultante.

El proceso de extracción de datos relacionales, tanto del EBD, IBD y de las RIs, se realiza mediante ERs. Las ERs se emplean para identificar relaciones y para identificar y validar atributos. Uno de los retos fue la identificación, validación y manipulación de las instancias de los atributos, debido a la gran variedad de formas en la cual pueden expresarse.

4.2 Integración de MOSTO

El prototipo para la generación de TMs fue integrado a una arquitectura con la finalidad de integrar herramientas para la gestión y validación de TMs (aplicación web). En este contexto, la Figura 4.3 presenta la arquitectura propuesta, que tiene una estructura basada en capas (datos, lógica y presentación) donde cada nivel cumple con un objetivo específico, lo que le permite ser escalable.

En los siguientes apartados se describe el funcionamiento de cada capa de la arquitectura propuesta para la gestión y validación de TMs.

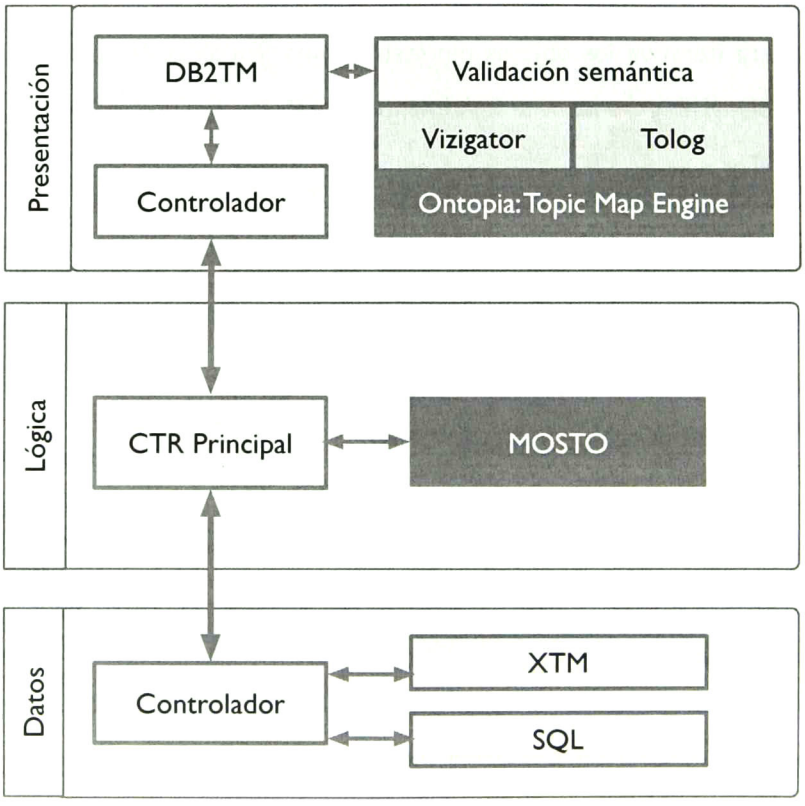


Figura 4.3: Arquitectura por capas para la integración del modelo propuesto MOSTO

4.2.1 Capa de presentación

La capa de presentación tiene tres elementos principales: un controlador, un módulo denominado DB2TM y un componente de validación integrado por cuatro partes. A continuación se describe cada uno de ellos:

Controlador, el componente tiene dos principales funciones de gestión. Primero, atender las solicitudes realizadas por la capa de presentación y enviarlas a la capa de lógica. La segunda función es atender las peticiones procedentes de la capa de lógica y distribuir las tareas a los componentes correspondientes en la capa de presentación.

DB2TM, su función principal es la gestión de los TMs. Algunas de sus tareas son: solicitar una *Base de Datos Relacional* (BDR) y enviar peticiones para su transformación a un TM.

Esquema de Validacion, que se compone de:

Ontopia Topic Map Engine: Es un motor encargado de gestionar la edición, navegación y visualización de los TMs.

Vizigator: Componente para la navegación y visualización de TMs. La visualización corrobora que el TM obtenido cumple los lineamientos del estándar ISO MDTM.

Tolog: Componente que integra un motor de consultas para TM.

Validación semántica: Módulo que representa un esquema de validación semántica de los TMs, esto empleando reglas de inferencia desarrolladas en el lenguaje *Tolog* [Garshol, 2006].

4.2.2 Capa de lógica

El módulo del prototipo se encuentra ubicado en la capa de lógica. Los elementos principales de la capa son:

Controlador el componente tiene tres funciones principales. Primero, atender las solicitudes realizadas por el componente MOSTO y enviarlas a la capa de presentación o de datos. La segunda función es atender las peticiones procedentes de la capa de presentación y distribuir las tareas a los componentes correspondientes. La tercera función es atender las peticiones procedentes de la capa de datos y distribuir las tareas a los componentes correspondientes (por lo regular a la capa de presentación).

MOSTO componente que representa la implementación de la metodología propuesta, descrita en la Sección 4.1.

4.2.3 Capa de datos

La capa de datos o persistencia está compuesta de tres elementos principales:

Controlador el componente tiene dos principales funciones de gestión. Primero, atender las solicitudes realizadas por capa de lógica. La segunda función es atender las peticiones de la misma capa, es decir, su interacción con los módulos XTM y SQL.

XTM es un componente encargado de la persistencia de los TMs.

SQL es un componente encargado de la persistencia de las BDRs.

5

Resultados

En este capítulo se presentan los experimentos y resultados obtenidos de la puesta en práctica de la metodología descrita en el capítulo anterior.

El método de evaluación sobre los *Topic Maps* (TMs) se centra en tres puntos principales. Primero se presenta un estudio del volumen de datos generado (TMs) por el enfoque propuesto en comparación con el volumen de datos de entrada (BDRs). Además se presenta una comparativa del volumen de representación y del tiempo de cómputo entre el enfoque propuesto y el enfoque más representativo del estado del arte presentado por Eslami *et al.* Segundo, se presenta la descripción de la validación sintáctica de los TMs resultantes de acuerdo al *Modelo de Datos de Topic Maps* (MDTM). Por último, se presenta la validación del aseguramiento de la calidad de las relaciones semánticas de los datos.

5.1 Banco de bases de datos relacionales

Los experimentos se realizaron a partir de un *benchmark* compuesto por 15 *Bases de Datos Relacionales* (BDRs) provenientes de distintos *Sistemas Manejadores de Bases de Datos* (SMBDs):

Microsoft® SQL Server, Oracle® y MySQL. La Tabla 5.1 muestra las características del *benchmark*[1] propuesto. Se indica el nombre de la BDR, el SMBD al que pertenece. También se muestra información con respecto al volumen de datos, como el número total de relaciones (tablas), el total de atributos y el total de instancias involucradas en cada BDR (incluye valores nulos y campos vacíos). Por último se muestra la etiqueta *real* o *ficticia* dependiendo de la procedencia de la BDR.

Tabla 5.1: Descripción del *benchmark* de BDRs propuesto para el proceso de experimentación

BDRs	SMBD	TIPO	RELACIONES	ATRIBUTOS	INSTANCIAS
Sales	SQL Server	Ficticia	4	15	39
Northwind	SQL Server	Ficticia	8	77	24275
Publications	SQL Server	Ficticia	11	64	1452
Adventure Works	SQL Server	Ficticia	10	96	37140
Human Resources	Oracle	Ficticia	7	35	1632
Order Entry	Oracle	Ficticia	8	52	12548
Mondial	Oracle	Real	33	134	95350
Classic Cars	MySQL	Ficticia	8	59	21179
Employees	MySQL	Real	6	24	132198
Hospital	MySQL	Ficticia	23	132	3855
Northwind	MySQL	Ficticia	13	88	24982
Presidents	MySQL	Ficticia	6	31	2202
Sakila	MySQL	Real	15	86	54801
Social Network	MySQL	Real	10	65	25482
World	MySQL	Real	3	24	27906

Cada elemento del *benchmark* de BDRs posee distintas características (descritas en la Tabla 5.1). Además de distintas restricciones de referencia y de dominio en su esquema que hacen compleja la tarea de representación de información en el estándar TM. A continuación se presenta una breve descripción de cada elemento del *benchmark* propuesto:

Sales: BDR ficticia sobre un sistema de ventas (utilizada para la demostración de la metodología propuesta).

Northwind: BDR ficticia sobre un sistema de inventarios y órdenes de compra. Ampliamente utilizada en el campo de las BDRs debido a que es apropiada para el aprendizaje sobre la

[1]Disponible en: http://www.tamps.cinvestav.mx/~ajose/mosto/benchmark.html

gestión de BDRs. Proporcionada por Microsoft para su producto Microsoft® SQL Server 2000.

Publications: BDR ficticia sobre un sistema de ventas y apartado de libros. Proporcionada por Microsoft® para su producto Microsoft SQL Server 2000.

Adventure Works: BDR ficticia de una empresa multinacional dedicada a la fabricación y venta de bicicletas. Proporcionada por Microsoft® para su producto Microsoft SQL Server 2008.

Human Resources: BDR ficticia sobre los recursos humanos de una empresa. Forma parte del conjunto de BDRs de prueba proporcionadas por Oracle®.

Order Entry: BDR ficticia sobre el seguimiento de pedidos, inventarios y venta de productos a través de diversos canales. Forma parte del conjunto de base de datos de prueba proporcionadas por Oracle®.

Mondial: BDR con datos geográficos; es una recopilación de distintas fuentes web, ha sido empleada como caso de estudio para la extracción e integración de información.

Classic Cars: BDR ficticia sobre una tienda de maquetas de coches clásicos. Fue proporcionada por Eclipse para las pruebas y experimentos de sus productos.

Employees: BDR ficticia sobre la gestión de usuarios y pagos de nómina en una compañía. La base de datos es proporcionada por MySQL para propósitos de prueba no triviales debido al gran volumen de datos.

Hospital: BDR que modela las entidades funcionales de un hospital. Se caracteriza por tener un gran número de relaciones y restricciones entre sus relaciones y datos.

Northwind: BDR ficticia sobre un sistema de inventarios y órdenes de compra. Inicialmente fue proporcionada por Microsoft para Microsoft SQL Server. Esta versión se encuentra también en MySQL.

Presidents: BDR ficticia proporcionada Paul DuBois como material adicional de su libro MySQL.

Sakila: BDR desarrollada por Mike Hillyer, proporciona un esquema estándar que es utilizado como ejemplo en los libros, tutoriales, artículos, etc. La BDR Sakila describe las principales características de MySQL tales como vistas, procedimientos almacenados y *triggers*.

Social Network: BDR real sobre la gestión de una red social. Ésta ha sido obtenida a partir de un proyecto de desarrollo tecnológico.

World: Esta BDR proporciona un conjunto de relaciones que contienen información sobre los países y ciudades del mundo.

5.1.1 Infraestructura de prueba

Las características del hardware empleado para la realización de los experimentos diseñados son los siguientes:

- Procesador Intel® Core2Duo a 2.93 GHz
- Memoria RAM de 8 GB
- Disco Duro de 500 GB

5.2 Generación de Topic Maps

El prototipo recibe como entrada la BDR compuesta por su *Esquema de Base de Datos* (EBD) y su *Estado de Base de Datos* (IBD). El proceso de extracción, manipulación y validación de datos se realiza mediante el uso de ER. La Figura 5.1 presenta el diagrama de generación automática de TMs basado en reglas de aprendizaje.

La evaluación del volumen de información semántica es la sumatoria del número de tópicos, asociaciones y ocurrencias (TAOs) obtenidos al procesar las BDRs. La Tabla 5.2 presenta las métricas TAO de los TMs resultantes al aplicar el enfoque propuesto.

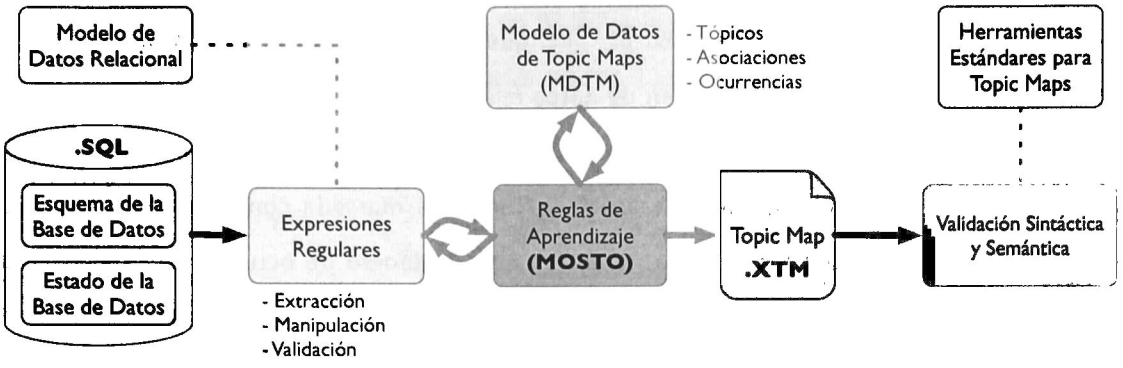


Figura 5.1: Esquema general para la obtención automática de TMs

Tabla 5.2: Descripción de los TMs resultantes al aplicar el enfoque propuesto MOSTO

TM	TOPICOS	ASOCIACIONES	OCURRENCIAS	TAOs	TIEMPO (Mins)
Sales	54	70	2	126	0.0005
Northwind	7128	27496	14	34638	16.3300
Publications	926	2081	0	3007	0.0756
Adventure Works	14821	42674	847	58342	63.4724
Human Resources	1030	2338	0	3368	0.1312
Order Entry	4617	14833	605	20055	7.943
Mondial	55516	110141	0	165657	362.4308
Classic Cars	5501	25527	23	31051	17.3952
Employees	44528	168293	0	212821	771.5921
Hospital	1884	5405	0	7289	0.7442
Northwind	5199	28344	9	33552	20.8347
Presidents	1415	3167	80	4662	0.2592
Sakila	12875	58387	600	71862	77.8831
Social Network	12650	30158	208	43016	26.4641
World	17189	40622	0	57811	41.4096

El número total de elementos TAOs en un TM, que representan la semántica de una BDR, es un indicador de la complejidad del volumen de datos relacionales. La Figura 5.2 muestra el total de tópicos, asociaciones, ocurrencias y la sumatoria de todas ellas (TAOs) para cada TM resultante. El experimento muestra que el número de asociaciones (línea marcada con un rombo) siempre es mayor para todas las instancias de prueba, mientras que el número de ocurrencias (línea marcada con un cuadro) representa el menor número de elementos en el TM. Por último, el número total de elementos TAOs se representa mediante la línea punteada marcada con una cruz.

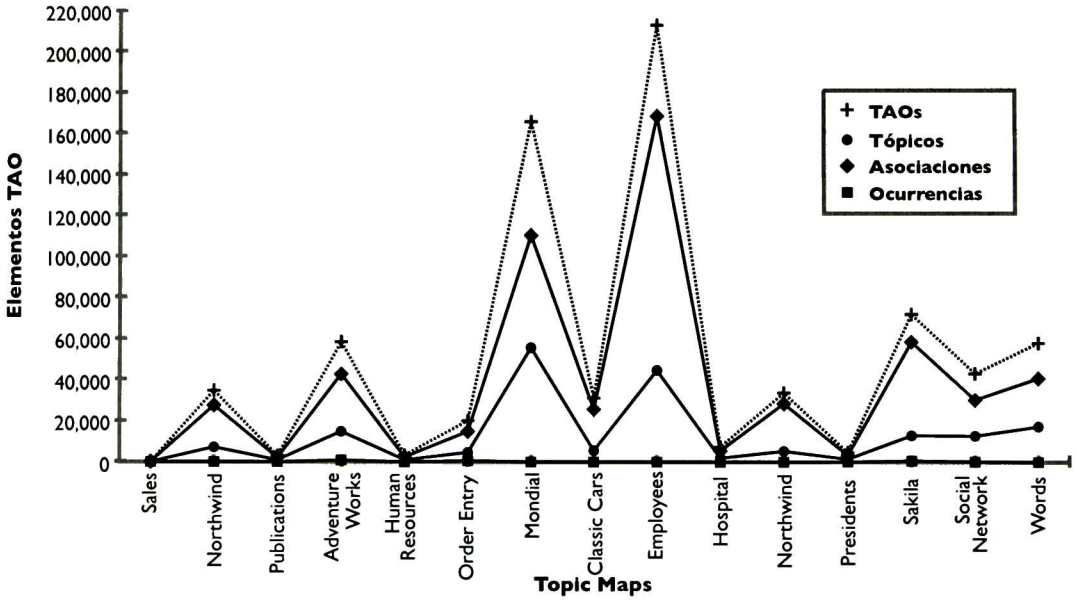


Figura 5.2: Comparativa de número de elementos TAOs

La Figura 5.3 muestra los resultados de la evaluación del rendimiento de los tiempos de cómputo empleado para la generación de los TMs. Se presenta una comparativa en porcentajes de tiempo de cómputo entre los elementos TAOs para cada TM generado. Por otro lado, la Figura 5.4 presenta una comparativa de tiempos promedio de cómputo del procesamiento de tópicos, asociaciones y ocurrencias de las 15 instancias de prueba. La Figura 5.4 muestra también que el tiempo porcentual promedio respecto al tiempo total empleado para el procesamiento de tópicos es de 9 %, de las asociaciones es de 88 % y de ocurrencias de un 3 %. Por lo tanto, el procesamiento de asociaciones

requiere de un tiempo porcentual promedio de 88 %, lo que significa que su procesamiento requiere el mayor tiempo de cómputo para la generación de los TMs.

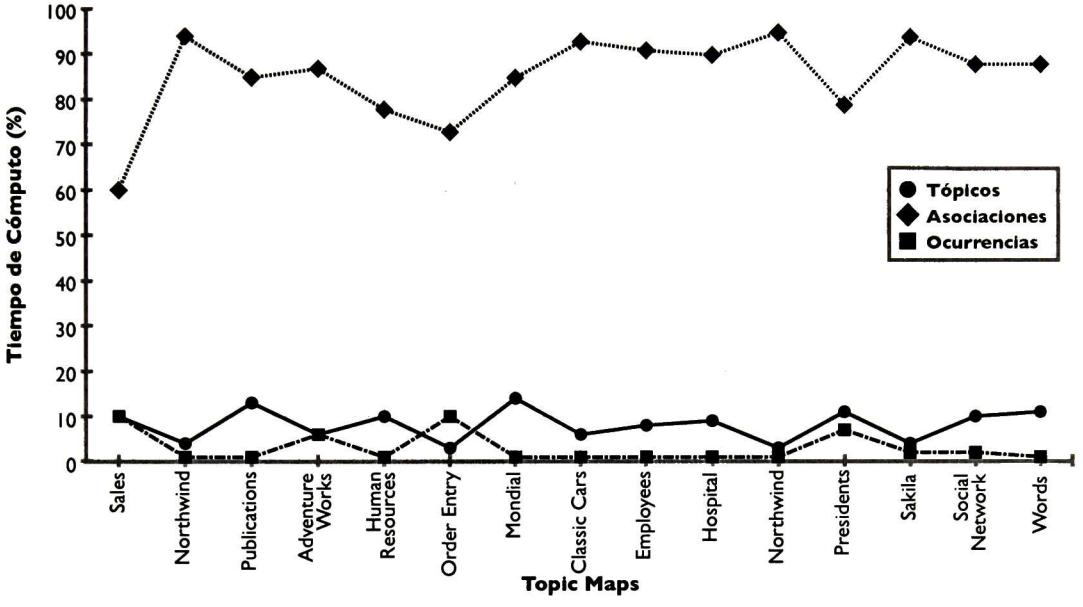


Figura 5.3: Comparativa del tiempo de cómputo porcentual para la generación de elementos TAOs

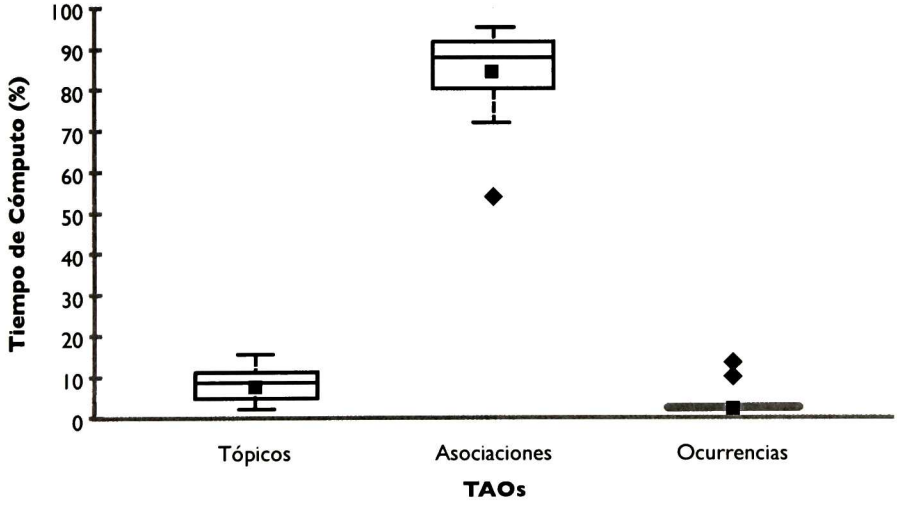


Figura 5.4: Comparativa del tiempo de cómputo porcentual promedio para la generación de los elementos TAOs

5.3 Comparativa del volumen de datos

En este apartado se presenta una comparativa acerca del volumen de datos generado (TM) por el enfoque propuesto en comparación con el volumen de datos de entrada (BDR). La Figura 5.5 presenta los resultados de la comparativa entre el volumen de datos de entrada (línea punteada marcada con un círculo) y el volumen de datos de salida (línea de trazos y puntos marcada con un cuadrado).

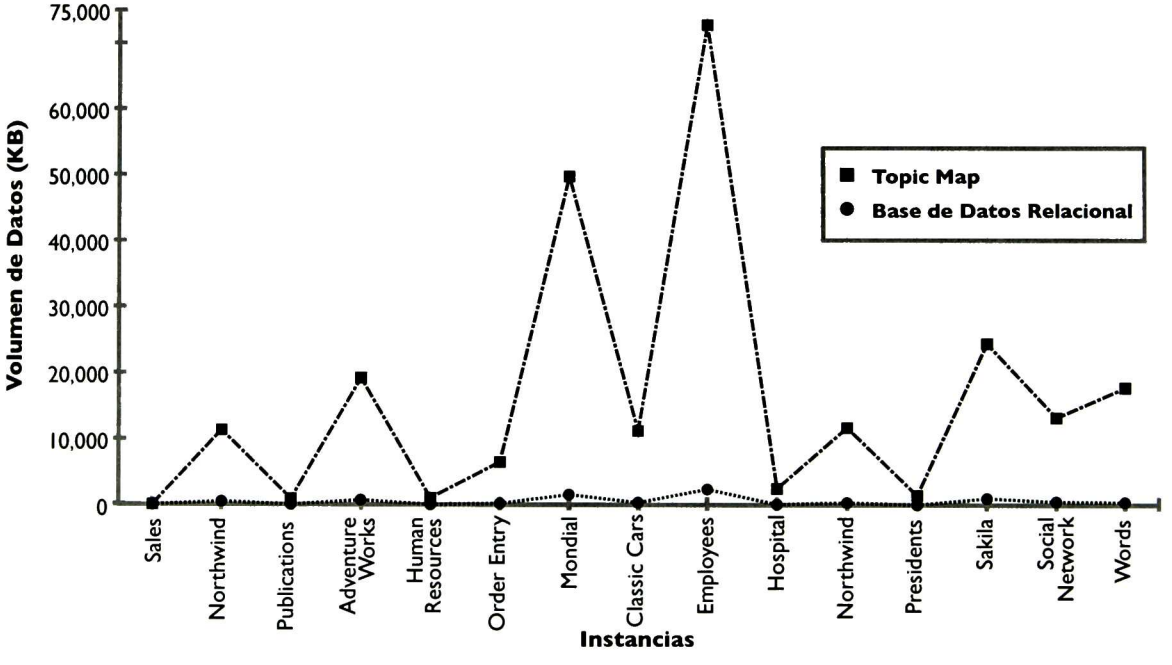


Figura 5.5: Comparativa del volumen de datos de entrada (BDR) y de salida (TM)

5.3.1 Comparativa del volumen de información

A continuación se presenta una comparativa entre el enfoque propuesto (MOSTO) y el enfoque más representativo del estado del arte propuesto por Eslami *et al.* El experimento incluye las siguientes comparativas (ver Tabla 5.3):

Tabla 5.3: Descripción de los TMs resultantes a partir del *benchmark* de BDRs

TOPIC MAP	MOSTO			ESLAMI ET AL.		
	TIEMPO (Mins)	TAOs	VOLUMEN (KB)	TIEMPO (Mins)	TAOs	VOLUMEN (KB)
Sales	0.0005	126	37	0.047	131	37
Northwind	16.3300	34638	11264	35.7888	60320	18330
Publications	0.0756	3007	893	1.005	4102	1243
Adventure Works	63.4724	58342	19148	78.492	76322	20741
Human Resources	0.1312	3368	1024	0.8286	3865	1229
Order Entry	7.943	20055	6451	13.5196	33079	10035
Mondial	362.4308	165657	49664	381.930	243675	65642
Classic Cars	17.3952	31051	11161	32.454	49287	16282
Employees	771.5921	212821	72704	811.927	318212	93047
Hospital	0.7442	7289	2457	1.2187	10704	3379
Northwind	20.8347	33552	11673	49.1950	59773	18944
Presidents	0.2592	4662	1433	0.1837	6301	1843
Sakila	77.8831	71862	24371	140.5775	107619	34611
Social Network	26.4641	43016	13209	52.8470	72934	21504
World	41.4096	57811	17715	63.4635	79466	23552

- Tiempo de cómputo empleado para el procesamiento de cada instancia de prueba (ver Figura 5.6).
- Número total de elementos TAOs (elementos principales del MDTMs) para cada instancia de prueba (ver Figura 5.7).
- Volumen de datos generado para la representación semántica de las instancias de prueba (ver Figura 5.8).

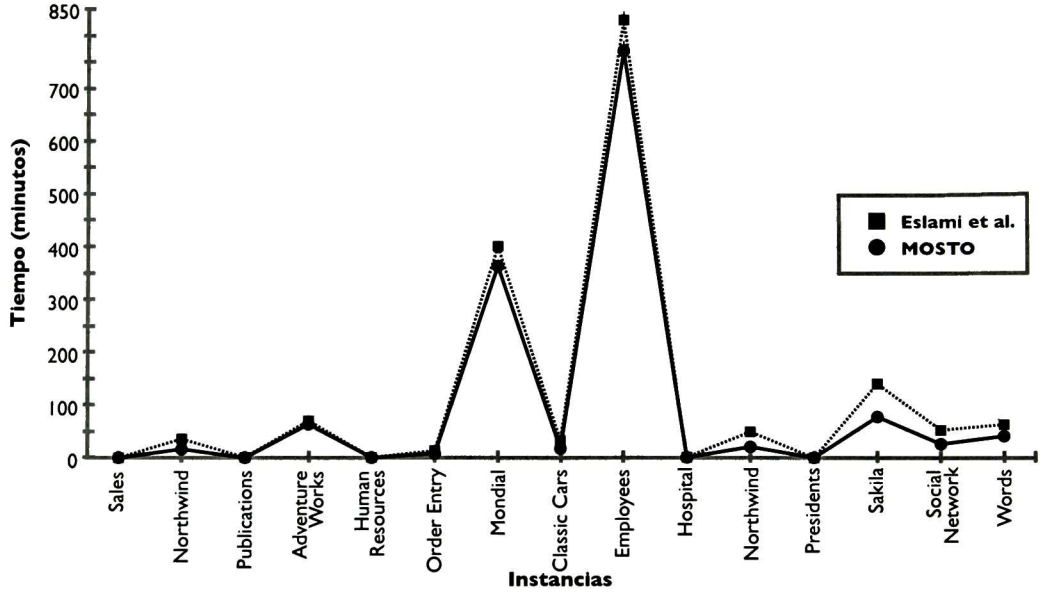


Figura 5.6: Comparativa del tiempo de cómputo entre el enfoque propuesto MOSTO y el propuesto por Eslami *et al.*

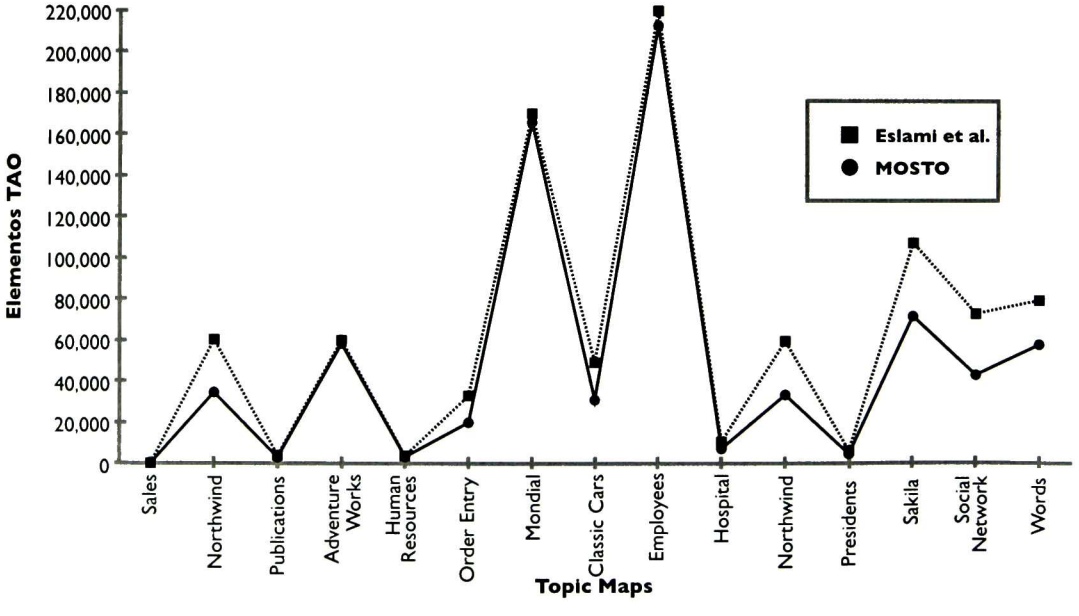


Figura 5.7: Comparativa del volumen de representación (elementos TAOs) entre el enfoque propuesto MOSTO y el propuesto por Eslami *et al.*

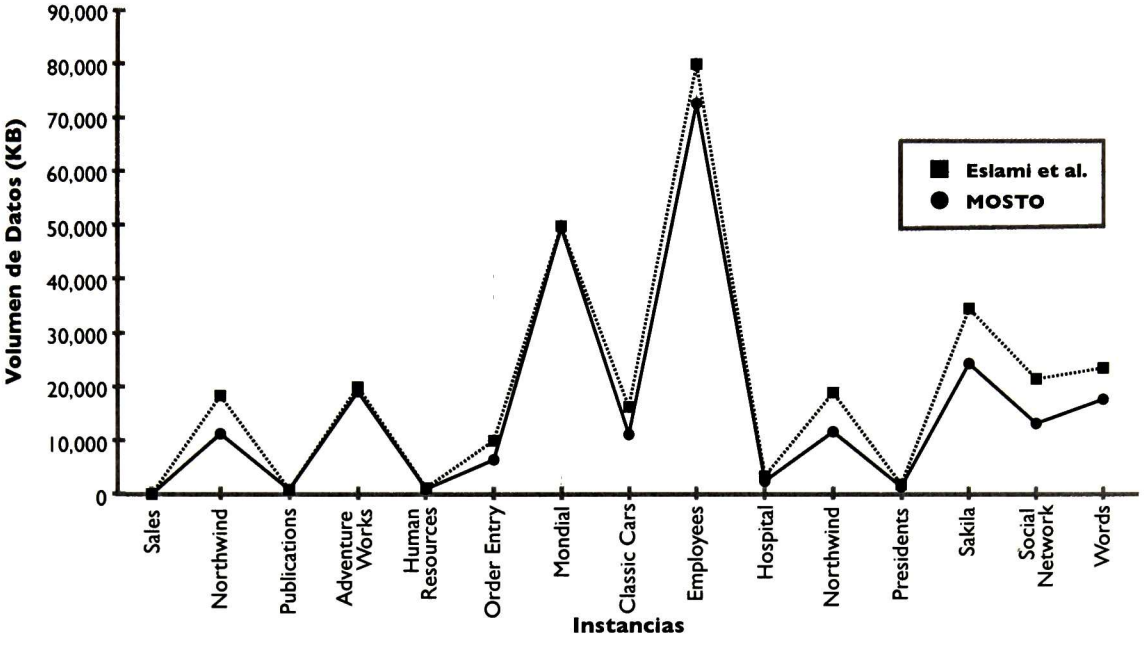


Figura 5.8: Comparativa del volumen de datos (KBs) entre el enfoque propuesto MOSTO y el propuesto por Eslami *et al.*

5.3.2 Análisis de resultados

A continuación se presenta un análisis de los resultados obtenidos tomando como base las características de cada método. Se presenta una tabla comparativa de características entre el enfoque MOSTO y el propuesto por Eslami *et al.*:

DESCRIPCIÓN	MOSTO	ESLAMI *ET AL.*
Esquema de la Base de Datos (EBD)	Si	Si
Estado de la Base de Datos (IBD)	Si	Si
Restricciones de Integridad (RIs)	Si	No
Redundancia de Datos	Si	No
Validación de valores nulos y campos vacíos	Si	*[2]
Proceso de transformación automático	Si	No

La comparativa puede considerarse justa debido a que ambos métodos emplean procesos similares para el proceso de transformación:

Similitudes: La principal similitud radica en que ambos métodos están basados en reglas. Además, ambos métodos consideran como entrada al proceso de transformación tanto al EBD como al IBD.

Diferencias. A continuación se listan las principales diferencias entre las metodologías en cuestión:

- El enfoque propuesto MOSTO se considera un enfoque automático, mientras que el enfoque propuesto por Eslami *et al.* es una metodología semi-automática. Se considera automático debido a que durante el proceso de transformación de los datos no interviene información externa. Por ejemplo, el enfoque de Eslami *et al.* requiere de un archivo de configuración que contiene información sobre los atributos que deben modelarse como tipos de ocurrencias. En MOSTO se emplean expresiones regulares para buscar, validar y definir los tipos de ocurrencias (ver Algoritmo 1).
- Ambos enfoques están basados en reglas. No obstante, la definición de cada regla es diferente. Por ejemplo, la metodología propuesta por Eslami *et al.* no proporciona información sobre como modelar las RIs de referencia cuando son del tipo *muchos-a-muchos*. MOSTO propone la creación de tópicos maestros que permiten asociar los datos y al mismo tiempo mantener la semántica entre ellos. Además, para cada regla se definen restricciones que ayudan a guiar de mejor manera el proceso de transformación.
- El enfoque propuesto MOSTO a diferencia de la metodología planteada por Eslami *et al.*, considera las distintas variantes de *Restricciones de Integridad* (RIs). Es importante considerarlas y modelarlas debido a que ayudan en la definición de la semántica de los datos. Por ejemplo, una persona debe tener asociada una edad que es representada por un número entero positivo. Es necesario realizar este tipo de validaciones porque semánticamente una persona no puede tener una edad negativa.

- El enfoque propuesto MOSTO a diferencia de la metodología propuesta por Eslami *et al.*, plantea un esquema de validación para la redundancia de tópicos. Debido a la naturaleza de los TMs donde los tópicos se encuentran comunicados mediante enlaces semánticos (asociaciones) no es necesario crear más de una instancia. Por ejemplo, supongamos que una BDR contiene un atributo que representa la edad de una persona, es muy probable que existan muchas personas con la misma edad y por lo tanto ese dato se encuentra repetido. El proceso de transformación evita crear tópicos repetidos, de tal forma que sólo se crean asociaciones entre los tópicos.

Por otra parte, se presento una comparativa entre el enfoque propuesto MOSTO y el enfoque propuesto por Eslami *et al.*. La comparativa se enfoco en dos puntos principales: (a) la medición el volumen de información generado por ambas propuestas y (b) la medición del tiempo computacional empleado para la transformación de los datos. Primero, se mostró que el enfoque propuesto produce menor volumen de información (ver Figura 5.7 y 5.8). Esto es debido a el enfoque contempla un proceso de validación para evitar la redundancia de datos. Por otro lado, el enfoque propuesto MOSTO requirió de menor tiempo de computo, a pesar de realizar una exploración de los datos relacionales para determinar los tipos de ocurrencias. Requiere de menor tiempo computacional por que además de evitar modelar datos redundantes también evita modelar sus respectivas asociaciones. En este contexto, la Figura 5.4 muestra que modelar las asociaciones consume el mayor tiempo computacional durante el proceso de construcción del TM.

5.4 Validación sintáctica de Topic Maps

Los TMs representan una de las principales propuestas para la visualización de información de la *Web Semántica* (WS). Existe una problemática obvia en este tema cuando se tiene en cuenta que un TM puede tener cientos de miles de asociaciones semánticas de diferente tipología [Grand and Soto, 2002].

Un TM se considera que está bien representado sintácticamente si cumple con los lineamientos definidos en estándar ISO/IEC 13250-7: *Notación Gráfica para Topic Maps* (NGTM) [ISO/IEC, 2007]. Por lo tanto, un TM generado por el enfoque propuesto es sintácticamente correcto sí y sólo sí es posible visualizarlo con una herramienta estándar reportada en el estado del arte. En este sentido, Ontopia Vizigator es un visualizador de TMs que cumple con el estándar ISO/IEC 13250-7, además de ser ampliamente utilizado por la comunidad de Topic Maps.

5.4.1 Visualización de asociaciones

Se ha mencionado que los principales elementos del MDTM son los tópicos, asociaciones y las ocurrencias. Los tópicos (elemento principal) poseen tres principales características: un nombre, ocurrencias y desempeña un determinado rol en una asociación.

A continuación se muestra la visualización de un TM desde la perspectiva de los tipos de asociaciones propuestos:

- Visualización del tipo de asociación *es-atributo-de* a partir de una relación (ver Figura 5.9).

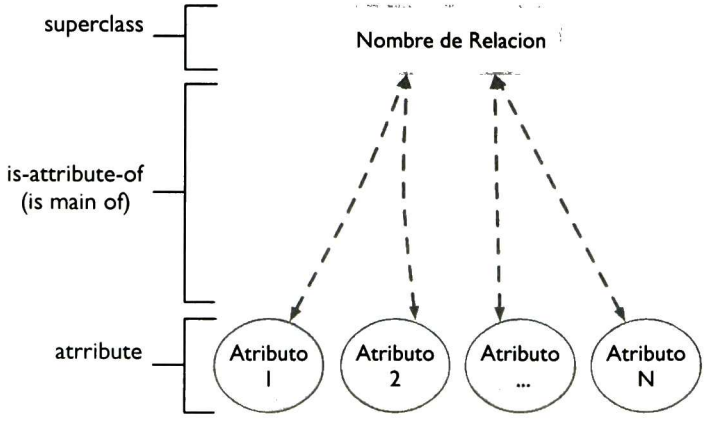


Figura 5.9: Visualización de las instancias del tipo de asociación *es-atributo-de* (is-attribute-of)

- Visualización del tipo de asociación *es-instancia-de* a partir de los atributos y sus respectivas instancias para una determinada relación (ver Figura 5.10).

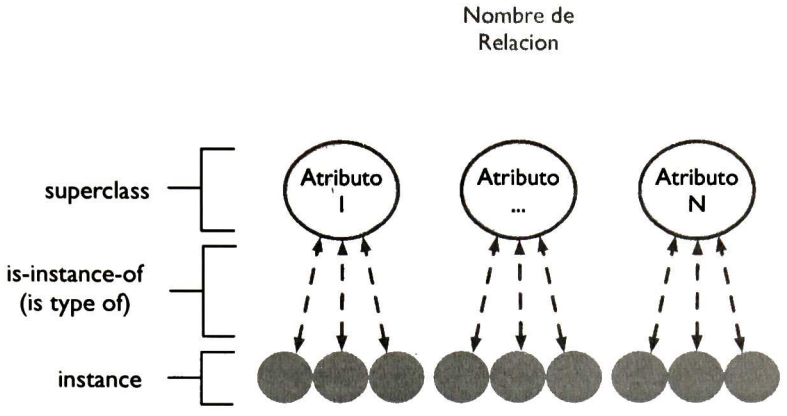


Figura 5.10: Visualización de las instancias del tipo de asociación *es-instancia-de* (is-instance-of)

- Visualización del tipo de asociación *tiene-relación-con* a partir de las instancias de una relación, donde la asociación se establece entre las instancias de una tupla con su respectiva llave primaria (ver Figura 5.11).

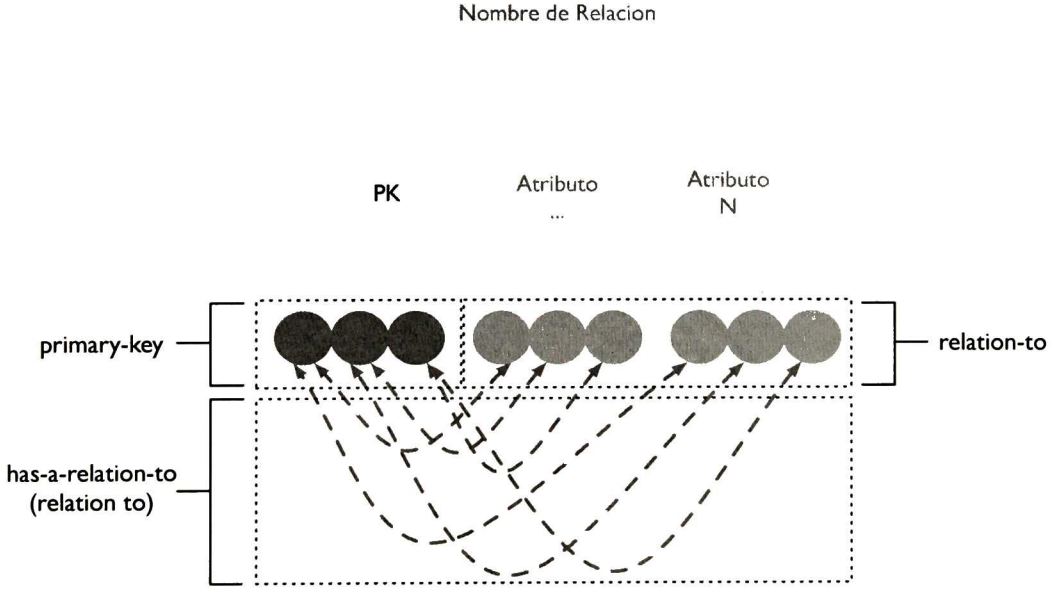


Figura 5.11: Visualización de las instancias del tipo de asociación *tiene-relacion-con* (has-a-relation-to)

- Visualización del tipo de asociación *tiene-referencia-con* a partir de las instancias de una relación, donde la asociación se establece a nivel de instancias entre los tópicos que representan la llave primaria y los tópicos que representan a la llave foránea (ver Figura 5.12).

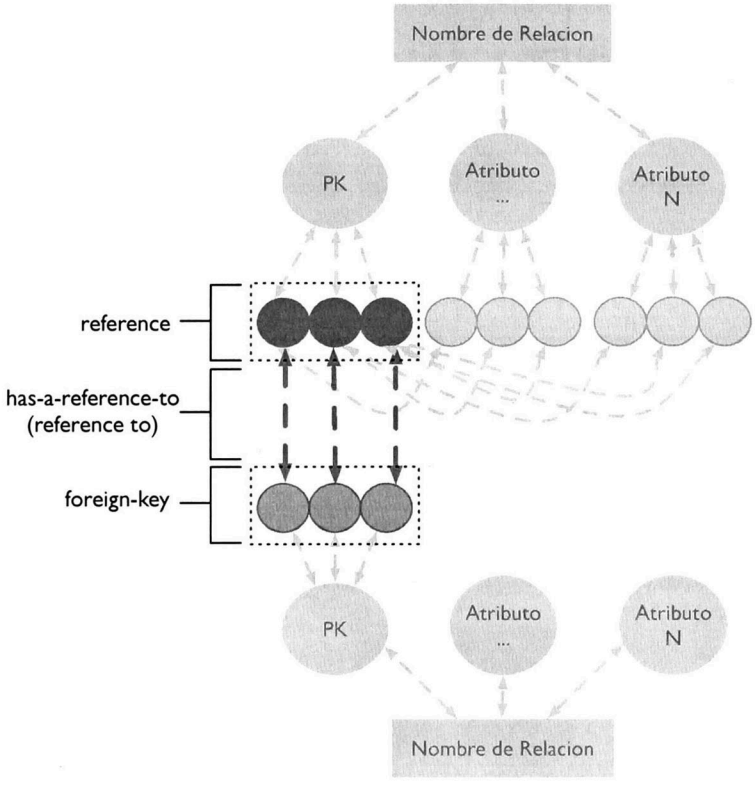


Figura 5.12: Visualización de las instancias del tipo de asociación *tiene-referencia-con* (has-a-reference-to)

5.4.2 Visualización con Vizigator

La validación sintáctica se realizó mediante la visualización de los TMs con la herramienta Vizigator [Ontopia, 2010]. A continuación se presenta la descripción del experimento utilizando el TM SALES, debido a que se trata de una instancia de prueba con pocos datos que facilitan su visualización y explicación.

La Figura 5.13 presenta la visualización del TM SALES creado con la herramienta Vizigator a partir del enfoque propuesto. De acuerdo a la figura, los tópicos son todas las entidades que pueden visualizarse (tópicos, tipos de tópicos e instancias). Los tipos de tópicos son presentados con rectángulos en color negro y sus instancias con rectángulos en color gris, mientras que las instancias que representan llaves primarias son presentadas con círculos en color gris. Los tópicos de referencia (tópicos temporales que representan las restricciones de integridad de referencia de una BDR) son representados con círculos en color negro. Por otro lado, las instancias de los tipos de asociaciones se representan mediante líneas sólidas y punteadas que relacionan a los tópicos. Las líneas sólidas en color gris son instancias del tipo de asociación *es-atributo-de*, las líneas punteadas en color gris representan a las instancias del tipo de asociación *es-instancia-de*, las líneas sólidas en color negro representan a las instancias del tipo de asociación *tiene-relación-con*. Por último, las instancias del tipo de asociación *tiene-referencia-con* se representan con las líneas punteadas en color negro.

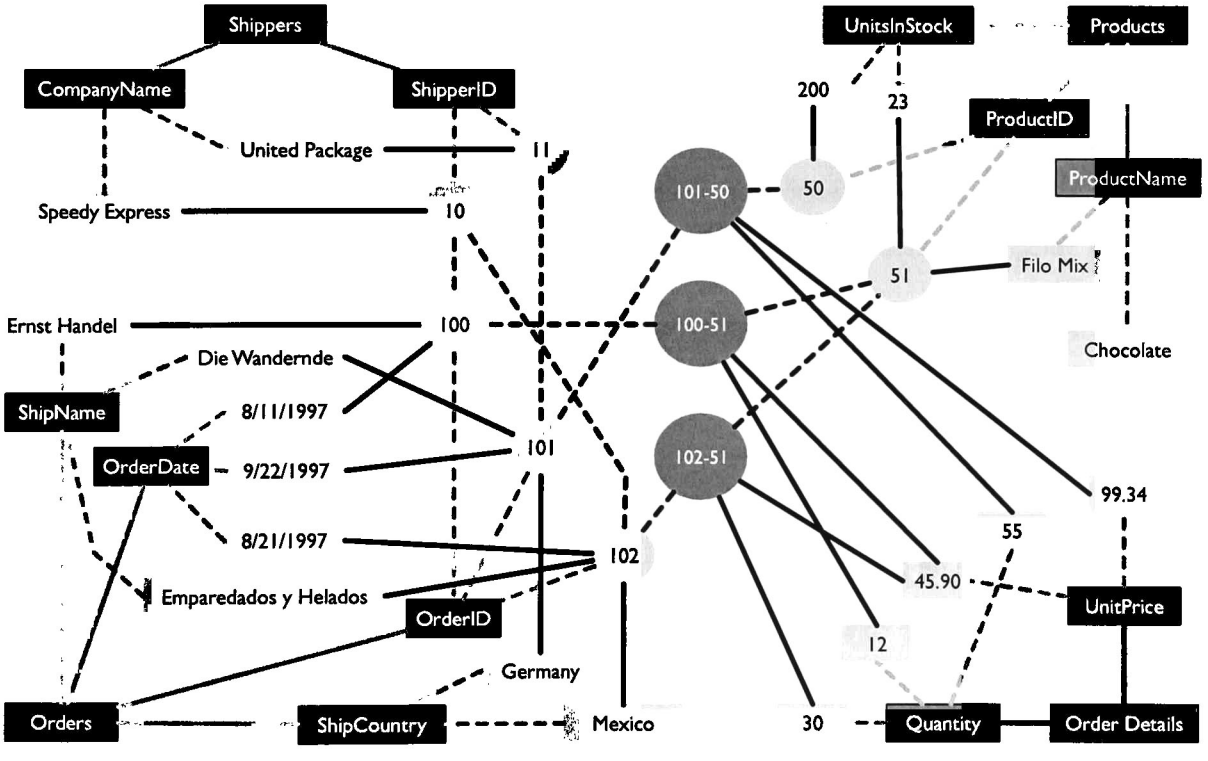


Figura 5.13: Visualización del TM SALES con la herramienta Vizigator

5.5 Validación semántica de Topic Maps

Se considera que un TM es semánticamente correcto si permite hacer inferencia a partir de la información que contiene. En este trabajo para hacer inferencia, se emplea el lenguaje de consultas Tolog [Garshol, 2006] el cual permite consultar los TMs obtenidos al aplicar el enfoque propuesto.

Tolog fue desarrollado por Ontopia para la recuperación de Topic Maps, inspirado por Datalog[3] (un subconjunto de Prolog) y SQL. Además, Tolog cumple con los requerimientos definidos en la norma ISO/IEC 18048:*Lenguaje de Consultas para Topic Maps* (LCTM) [ISO/IEC, 2004]. Algunas características soportadas por el lenguaje son funciones de agregación, de proyección, ordenación y paginación de resultados.

Un TM es una estructura abstracta que contiene información de interés sobre una BDR relacionada a una comunidad específica de usuarios. El objetivo de tener la información estructurada en una BDR es con la finalidad de facilitar el proceso de búsqueda de la información almacenada. En este sentido, Tolog permite la recuperación de información en Topic Maps similar al de una BDR con SQL. El lenguaje de consulta SQL se utiliza para definir el esquema de la BDR, actualizar y modificar la BDR. Estas operaciones también pueden ser aplicadas a un repositorio de Topic Maps mediante Tolog, en el caso del LCTM las tareas que Tolog es capaz de realizar son los siguientes:

- Consultas que retornan como resultado tópicos y TMs,
- Agregar información al TM en todos sus niveles de granularidad y
- Borrar información del TM en todos sus niveles de granularidad.

Tolog es un lenguaje de programación declarativo basado en la lógica de predicados de primer orden. Por lo tanto, mediante Tolog es posible definir un esquema de validación semántica para estudiar la inferencia de la información de los TMs generados a partir de las BDRs. En un Topic Map las declaraciones del universo del discurso definido por el dominio de la BDR se expresan mediante asociaciones.

[3]Es un lenguaje de consultas y reglas para bases de datos deductivas.

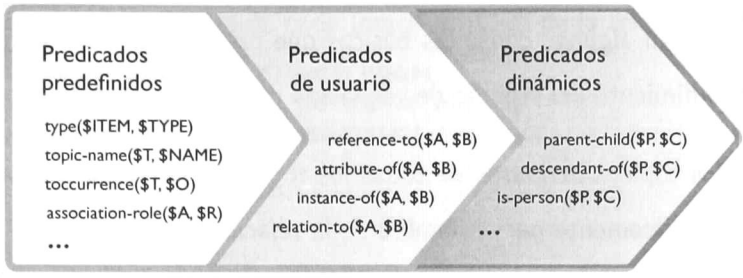


Figura 5.14: Tipos de predicados en Tolog

En Tolog existen tres tipos de predicados: a) los definidos por el usuario (a través de la declaraciones de asociaciones y ocurrencias), b) los predefinidos por el lenguaje (predicados basados en el LCTM) y c) los predicados dinámicos (creados a partir de tipos de ocurrencias y asociaciones en el TM). Los tipos de predicados mencionados y su secuencia de uso se muestran en la Figura 5.14.

5.5.1 Esquema de validación semántica

Los tipos de asociaciones y roles propuestos (*is-instance-of*, *is-attribute-of*, *has-a-relation-to* y *has-a-reference-to*) ayudan a construir un TM con enlaces semánticos entre los diferentes tópicos, estos tipos de asociaciones son nombrados como tipos de predicados durante el proceso de validación semántica [Gronmo, 2000]. La Figura 5.15 muestra los componentes de un predicado a partir del tipo de asociación *is-attribute-of*.

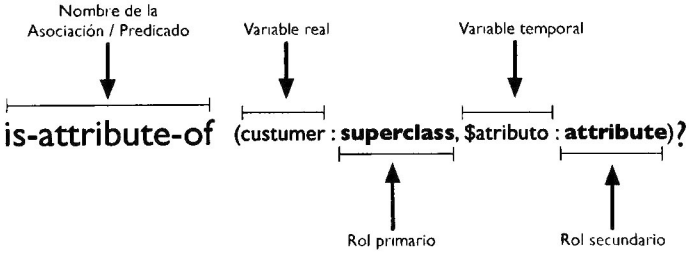


Figura 5.15: Componentes de una consulta básica con Tolog

Mediante el uso de los predicados predefinidos por Tolog y los predicados propuestos (obtenidos a partir de los tipos de asociaciones y tipos ocurrencias), es posible realizar consultas complejas que involucren distintos tipos de tópicos similares a las consultas que pueden realizarse con los SMBDs.

A continuación se listan algunas consultas básicas que forman parte del esquema de validación semántica para el entendimiento del proceso de validación de TM SALES mediante el lenguaje Tolog:

1. Obtener todos los tópicos *atributos* de todas los tópicos que representan *relaciones* (tablas), ordenados lexicográficamente por el nombre de la *relación*:

```
SELECT   $RELATION, $ATTRIBUTE FROM
         is-attribute-of($RELATION : superclass, $ATTRIBUTE : attribute)
ORDER BY $RELATION
?
```

2. Obtener todas las *instancias* de todos los tópicos *atributos* de cada *relación*, ordenados lexicográficamente por el nombre de la relación y atributos.

```
SELECT   $RELATION, $ATTRIBUTE, $INSTANCE FROM
         is-instance-of($ATTRIBUTE : class, $INSTANCE : instance),
         is-attribute-of($RELATION : superclass, $ATTRIBUTE : attribute)
ORDER BY $RELATION, $ATTRIBUTE
?
```

3. Obtener todas las instancias de todas los tópicos *llaves primarias* y su respectivo tópico *atributo* y nombre de *relación*, ordenados lexicográficamente por *relación*, *atributos* y *llave primaria*.

```
SELECT   $RELATION, $ATTRIBUTE, $PK, $INSTANCE FROM
         has-a-relation-to($PK : primary-key, $INSTANCE : relation-to),
         is-instance-of($ATTRIBUTE : class, $PK : instance),
         is-attribute-of($RELATION : superclass, $ATTRIBUTE : attribute)
ORDER BY $RELATION, $ATTRIBUTE ,$PK
?
```

4. Obtener todas las *instancias* y *atributos* de la relación SHIPPERS.

```
SELECT   $ATTRIBUTE, $INSTANCE FROM
         is-attribute-of(shippers: superclass, $ATTRIBUTE : attribute),
         is-instance-of($ATTRIBUTE : class, $INSTANCE : instance)
ORDER BY $ATTRIBUTE
?
```

5. Obtener el número de instancias de cada tópico *atributo* de la relación ORDERS.

```
SELECT    $ATTRIBUTE COUNT($INSTANCE) FROM
          is-attribute-of(orders: superclass, $ATTRIBUTE : attribute),
          is-instance-of($ATTRIBUTE : class, $INSTANCE : instance)
ORDER BY  $ATTRIBUTE
?
```

6. Obtener los nombres de los países que ha atendido la compañía *"Speed Express"* (esta consulta requiere el uso de las relaciones SHIPPERS y ORDERS).

```
SELECT    $COUNTRY FROM
          has-a-relation-to($SHIPPER_PK : primary-key, idcompanynamespeedexpress: relation-to),
          has-a-reference-to($SHIPPER_PK: reference, $ORDERS_PK : foreign-key),
          has-a-relation-to($ORDERS_PK : primary-key, $COUNTRY: relation-to),
          is-instance-of($COUNTRY : instance, $shipcountry-order : class)
ORDER BY  $ATTRIBUTE
?
```

7. Obtener los nombres de los países que NO ha atendido la compañía *"Speed Express"* y que hayan atendido otras compañías.

```
SELECT    $COUNTRY FROM
          NOT (has-a-relation-to($SHIPPER_PK : primary-key, idcompanynamespeedexpress: relation-to),
          has-a-reference-to($SHIPPER_PK: reference, $ORDERS_PK : foreign-key)),
          has-a-relation-to($ORDERS_PK : primary-key, $COUNTRY: relation-to),
          is-instance-of($COUNTRY : instance, $shipcountry-order : class)
ORDER BY  $ATTRIBUTE
?
```

8. Obtener los nombres de los clientes (ShipName) y la fecha, que ha atendido cada compañía (esta consulta requiere el uso de las relaciones SHIPPERS y ORDERS).

```
SELECT    $COMPANY, $SHIPNAME, $DATE FROM
          is-instance-of($COMPANY : instance, companyname-shippers : class)
          has-a-relation-to($COMPANY : relation-to, $SHIPERS-PK: primary-key),
          has-a-reference-to($SHIPPER_PK: reference, $ORDERS_PK : foreign-key),
          is-instance-of($DATE : instance, orderdate-orders : class)
          has-a-relation-to($ORDERS_PK : primary-key, $DATE: relation-to),
          is-instance-of($SHIPNAME : instance, shipname-orders : class)
          has-a-relation-to($ORDERS_PK : primary-key, $SHIPNAME: relation-to)
ORDER BY  $ATTRIBUTE ASC
?
```

9. Obtener los sitios web (*Occurrence*, Website) de todas las compañías.

```
SELECT   $PK, $COMPANY, $WEBSITE FROM
         is-instance-of(shipperid-shippers: class, $PK : instance)
         has-a-relation-to($PK : primary-key, $COMPANY: relation-to),
         is-instance-of($COMPANY: instance, companyname-shippers: class)
         website-shippers($PK: reference, $WEBSITE),
ORDER BY $COMPANY
?
```

10. Obtener un detallado de las órdenes de compra realizadas por el negocio, para lo cual se requiere conocer:

    a) El nombre del cliente que realizo el pedido.

    b) El nombre del producto solicitado.

    c) La cantidad de productos solicitados.

    d) El nombre de la compañía de envío de la orden.

```
SELECT   $SHIPNAME, $PRODUCT, $COMPANY, $UNITPRICE, $QUANTITY FROM
         is-instance-of($PRODUCT : instance, productname-products : class)
         has-a-relation-to($PRODUCT : relation-to, $PRODUCT-PK: primary-key),
         has-a-reference-to($PRODUCT-PK: reference, $MASTER-PK : foreign-key),
         has-a-relation-to($MASTER-PK : primary-key, $QUANTITY : relation-to),
         is-instance-of($QUANTITY : instance, quantity-orderdetails : class)
         has-a-relation-to($MASTER-PK : primary-key, $UNITPRICE : relation-to),
         is-instance-of($UNITPRICE : instance, unitprice-orderdetails : class)
         has-a-reference-to($MASTER-PK: foreign-key, $ORDERS-PK : reference),
         has-a-relation-to($ORDERS-PK : primary-key, $SHIPNAME: relation-to)
         is-instance-of($SHIPNAME : instance, shipname-orders : class)
         is-instance-of($COMPANY: instance, companyname-shippers: class)
         has-a-relation-to($COMPANY : relation-to, $SHIPERS-PK: primary-key),
         has-a-reference-to($SHIPPER_PK: reference, $ORDERS_PK : foreign-key),
ORDER BY $ATTRIBUTE ASC
```

5.5.2 Reglas de inferencia

Un aspecto interesante sobre la utilización de un TM es la inferencia de información. Es decir los mecanismos por los que podemos obtener información no explícita en el TM.

Las inferencias que se pueden realizar a partir de un conjunto de asociaciones están marcadas por la propiedades que tienen los tipos de asociaciones implicadas. Aunque en la literatura se mencionan distintas propiedades, en esta sección nos centraremos en la *transitiva* y *simétrica* por ser las más empleadas para inferencia y creación de TMs. A continuación se analizan estas propiedades:

Simetría Establecen una relación entre dos tipos de roles idénticos o similares sin modificar su relación semántica.

Transitividad Permiten aclarar implícitamente un hecho a través de varias asociaciones y se especifican mediante los tipos de asociación *superclase-subclase* (*is-attribute-of*). Por ejemplo, *María es una instancia de nombre* o *María es el valor de la etiqueta nombre* y *nombre es un atributo de persona*, se puede inferir que María es una persona.

En muchos casos existen relaciones implícitas en el TM que no fueron predefinidas como tipos de asociaciones pero que pueden deducirse a partir de relaciones más básicas, es decir, a partir de los tipos de asociaciones definidas de forma explícita. Mediante las reglas de inferencia es posible encontrar estas relaciones implícitas a través de la declaración de reglas simples. Posteriormente, la regla puede utilizarse en tareas posteriores para simplificar las consultas.

Las reglas de inferencia pueden también utilizar a otras reglas de inferencia, lo que significa que es posible crear grandes estructuras de razonamiento añadiendo capas de reglas de inferencia. Las reglas de inferencia también pueden llamarse así mismas, lo que hace posible definir reglas de inferencia recursivas.

Un ejemplo de una regla de inferencia para una relación genérica *padre-hijo*, podría ser la siguiente:

```
desendant-of($ANC, $DESC):- {
      parent-child ($ANC: parent, $DESC : child)|
      parent-child ($ANC: parent, $MID : child),
      descendant-of($MID, $DESC)
}.
```

Obsérvese cómo la regla es esencialmente una formalización de lo que significa ser un descendiente. Es decir, el ancestro es el padre del descendiente, o el ancestro es el padre del algún tópico intermedio del cual existe un descendiente, esto debido a la recursividad en el último paso de la regla de inferencia. A continuación se presentan las reglas de inferencia para obtener los tópicos que tienen una referencia al tópico *orders*:

```
1  parent-child($PARENT, $CHILD) :- {
2    attribute-of($PARENT : superclass, $CHILD : attribute) |
3    instance-of($CHILD, $PARENT)
4  }.
5
6  descendant-of($ANC, $DES) :- {
7    parent-child($ANC : parent, $DES : child) |
8    parent-child($ANC : parent, $MID : child), descendant-of($MID, $
9  }.
10
11 reference-to($PARENT : reference, orders : refers-to),
12 descendant-of($PARENT, $DESENDANT)?
```

5.5.3 Inferencia sobre el Topic Map *NORTHWIND*

La realización del experimento de validación semántica se realiza mediante el TM NORTHWIND obtenido a partir de una BDR sobre un sistema de información que registra las operaciones de gestión de un restaurante.

Se ha utilizado al TM NORTHWIND para revisar el tipo de conocimiento que reside en la BDR NORTHWIND. Se construyó un conjunto de consultas con distintas características. La Tabla 5.4 muestra 12 consultas construidas en Transact SQL y en el lenguaje de consultas Tolog para TM.

Tabla 5.4: Comparativa de consultas entre SQL y Tologa para la instancia NORTHWIND

Consulta ID	Descripción de la Consulta	SQL	Tolog
Q_01	Get the *products* for each *category*	True	True
Q_02	Get *persons* with an old of 10 years in the company and who are over 40 years.	False	True
Q_03	Get *orders* made in January 1997.	True	True
Q_04	Get *orders* with no shipping company or have been made in January.	True	True
Q_05	Get the last 10 *orders* in 1999.	True	True
Q_06	Get the 3 most expensive products each supplier.	False	True
Q_07	Get the number of *orders* for each *employee* made in 1998.	True	True
Q_08	Get the 3 best selling products.	False	True
Q_09	Get the sales amount per month of each year.	True	False
Q_10	Get the total sales for each employee performed by region.	True	False
Q_11	Get total territories served by each employee.	True	True
Q_12	Get products more expensive and sold.	False	True

De acuerdo al conjunto de consultas presentadas es posible concluir lo siguiente: a) la mayoría de las consultas resueltas por SQL también se pueden realizar con Tolog, b) las consultas Q_09 y Q_10 que no fueron resueltas por Tolog se debe a que requieren de distintos grupos y filtros entre las restricciones de integridad de referencia entre las relaciones (algo propio de SQL) y c) las consultas Q_02, Q_06 y Q_08 que no fueron resueltas por SQL es debido a que estas consultas incluyen información implícita que no se encuentra almacenada de manera explícita en la BDR original NORTHWIND. Esta información implícita fue obtenida mediante reglas de inferencia con el lenguaje Tolog de manera satisfactoria. La recuperación de información con reglas de inferencia demuestran que el enfoque propuesto es semánticamente correcto.

6

Conclusiones y Trabajo Futuro

El objetivo de este trabajo de investigación fue proponer una metodología basada en reglas de aprendizaje para la representación del conocimiento a partir de datos relacionales, esto mediante el uso del estándar *Topic Maps* (TMs).

El enfoque propuesto consiste en la obtención de TMs a partir de *Bases de Datos Relacionales* (BDRs). La solución propuesta, llamada MOSTO, demostró tener mejores resultados cuando se le comparó con el enfoque más representativo del estado del arte. También se mostró que MOSTO genera TMs válidos, esto debido a que fueron validados de manera sintáctica y semántica mediante herramientas estándares reportadas en el literatura especializada. Además, se propuso un esquema validación sobre la semántica de los TMs obtenidos donde se demostró la inferencia de información.

6.1 Conclusiones

Las siguientes conclusiones fueron obtenidas durante el proceso de elaboración de la metodología y de la validación experimental del enfoque propuesto:

- El uso de reglas de aprendizaje con restricciones ayudan a guiar de manera exitosa el proceso de transformación para la obtención de Topic Maps válidos a partir de datos relacionales.

- Además del Esquema y el Estado de la BDR fue necesario considerar las restricciones de integridad de referencia, de dominio y de rango de datos. Se observó que la incorporación de estas restricciones ayudan de manera significativa a la definición de la semántica de los datos.

  La incorporación de la validación del dominio y del rango de datos no incrementa de forma significativa el volumen de representación y de datos, mas sin embargo ayudan a definir de manera correcta la semántica de los datos involucrados.

- Se propuso un *benchmark* de 15 BDRs con la finalidad de probar la robustez y escalabilidad del enfoque propuesto. Los resultados obtenidos por nuestro enfoque a partir del *benchmark* muestran que la generación de asociaciones (relaciones semánticas entre tópicos) requiere en promedio de un 88 % del total del tiempo de cómputo para procesar la BDR, el resto del tiempo promedio, 22 %, se emplea en la generación de tópicos y ocurrencias.

  Con respecto al volumen de representación de los elementos TAO, el número de asociaciones crece de manera significativa con respecto al número de tópicos y ocurrencias.

  El volumen de datos es un factor importante a considerar debido a que indica en qué proporción se incrementa el volumen de datos de salida (representación del conocimiento) con base en el volumen de datos de entrada (BDR). En este sentido, el volumen de datos para la representación del conocimiento de una BDR mediante el estándar TM se incrementa en promedio 27 veces respecto al volumen original.

- Se realizó un comparativa entre el enfoque propuesto y el propuesto por Eslami *et al.*, en la cual se observó que nuestro enfoque obtuvo mejores resultados, tanto en la calidad de resultados, en el costo computacional y en el volumen de datos generado para la representación de la información a partir del mismo conjunto de instancias de prueba. El enfoque propuesto genera un volumen de datos menor que el enfoque propuesto por Eslami *et al.* para la representación

de la información. Esto debido a que nuestro enfoque incorpora un proceso de validación del dominio y del rango de los tipos de datos al transformar la BDR.

- Los tipos de asociaciones propuestas fueron utilizadas como predicados para la inferencia de información no explícita en los TMs.

6.2 Aportaciones

Las principales aportaciones científicas de este trabajo de tesis son:

- Un método basado en reglas para la obtención de TM a partir de BDR. Esta contribución fue publicada en:
  - Adan Jose-Garcia, Ivan Lopez-Arevalo, Victor Sosa-Sosa, "Building Topic Maps from Relational Databases". En: *International Conference on Electrical Engineering, Computing Science and automatic Control (CCE 2012)*, IEEE Press. Ciudad de México, 2012, pp. 294-299.
- Un esquema de validación sobre la semántica de los TMs obtenidos mediante reglas de inferencia.
- Un *benchmark* de prueba compuesto por BDRs tanto reales como ficticias.
- Una implementación de la metodología propuesta y su integración a una aplicación web para la gestión de TMs.
- Un documento de tesis que reporta el trabajo de investigación realizado.

6.3 Dificultades

Durante el desarrollo de la metodología propuesta se presentaron distintas dificultades. Pueden clasificarse en tres grupos principales:

Obtención de TMs. Uno de los principales retos en el proceso de transformación de la BDR fue la lectura, validación y manipulación de los datos relacionales, debido a las variantes en las cuales éstos pueden presentarse. Este problema fue solventado mediante el uso de expresiones regulares.

Seleccionar al mejor conjunto de tipos de asociaciones necesarias capaz de enlazar de manera semántica a los datos relacionales. Además, establecer las condiciones para tratar con las *Restricciones de Integridad* (RIs) de referencia de la BDR.

La proceso de identificación automático de tipos de ocurrencias fue una tarea difícil, debido a que se tiene que leer e interpretar cada instancia de cada atributo en la BDR para determinar si puede considerarse como un tipo de ocurrencia, por lo tanto es una tarea que además demanda un alto costo computacional.

Selección de herramientas. Una tarea difícil fue el proceso de validación y selección de un conjunto de tecnologías estándares reportadas en la literatura especializada para visualizar y consultar un TM. Primero, seleccionar una herramienta para visualizar un TM que cumpliera con el estándar ISO/IEC (NGTM) y además que permitiera navegar de manera gráfica entre los componentes semánticos del TM obtenido. Distintas herramientas fueron probadas, la herramienta *Ontopia Vizigator* fue la seleccionada. *Vizigator* es una herramienta que cumple con el estándar ISO/IEC (LCTM) para consultar al TM generado y así establecer un esquema de validación semántica de acuerdo a la metodología propuesta. El lenguaje seleccionado para hacer consultas e inferencias fue *Tolog*, debido a que se trata de un lenguaje lógico y declarativo, lo que permite trabajar con reglas de inferencia.

***Benchmark* de BDRs.** Buscar y seleccionar un conjunto de bases de datos representativas para probar el enfoque propuesto resultó una tarea difícil debido a que se buscaban BDRs de diferente volumen de datos, número de relaciones, atributos, instancias, restricciones de integridad, etc.

6.4 Trabajo Futuro

Con base al trabajo realizado, se considera que aún quedan algunos aspectos pendientes que podrían abordarse para mejorar el trabajo que se ha realizado. A continuación se describen las más importantes:

- Actualmente el enfoque propuesto asume que los nombres de los atributos son etiquetas válidas, es decir, están sintácticamente bien escritos y semánticamente bien contextualizados (tarea dentro del diseño de la BDR). Se realizó una validación sintáctica sobre el lenguaje que materializa a los TMs y también se realizó una validación semántica sobre las asociaciones que relacionan a los tópicos. No obstante, esto no garantiza que los nombres de los atributos sean valores de etiquetas válidos. Por lo tanto, se requiere de un modelo de representación adicional que permita unificar y generalizar el nombre de los atributos de acuerdo al dominio semántico de la BDR mediante la reutilización de ontologías que pertenezcan al dominio y del uso de jerarquías de conceptos para determinar el grado de similitud entre el valor del atributo y un concepto que pertenece a la jerarquía.

- También se propone realizar un modelo de estimación del tiempo de cómputo empleado para la obtención de los TMs. Esta estimación se puede realizar utilizando como base teórica el tiempo de cómputo empleado por el enfoque propuesto para el *benchmark* de BDRs propuesto. Esto debido a que el proceso de transformación no requiere de elementos estocásticos para la obtención del TM; es decir, se trata de un proceso determinista. Este modelo de estimación ayudaría a determinar de manera precisa el costo del tiempo computacional requerido para la generación de un TM a partir de cualquier BDR.

- Dadas las ventajas y características de los TMs, por ejemplo su facilidad para navegar, consultar e inferir información proporcionan una base para realizar búsquedas en lenguaje natural; por lo cual sería conveniente implementar una interface para este tipo de consultas con las adaptaciones necesarias para trasladar la consulta a la nomenclatura de los TMs.

Bibliografía

[Ahmed, 2009] Ahmed, K. (2009). Topic maps sparql. [Online: http://tmra.de/2009/talks/TMSPARQL, Visitado en: Julio-2012].

[Barta, 2004] Barta, R. (2004). Topic maps - data model. [Online: http://astma.it.bond.edu.au/querying.mhtml, Visitado en: Abril-2012].

[Bizer and Cyganiak, 2004] Bizer, C. and Cyganiak, R. (2004). D2r server – publishing relational databases on the semantic web. In *3rd International Semantic Web Conference*.

[Ching-Song Don Wei, 2011] Ching-Song Don Wei, Jiann-Gwo Doong, N. P. A. (2011). Learning semantics from relational database schema and data. *Energy Procedia*, 13:6257–6266.

[Davis et al., 1993] Davis, R., Shrobe, H., and Szolovits, P. (1993). What is a knowledge representation? *Artificial Intelligence Magazine*, 14(1):17–33.

[Dicheva and Dichev, 2006] Dicheva, D. and Dichev, C. (2006). Tm4l: Creating and browsing educational topic maps. In *British Journal of Educational Technology*, volume 3, pages 391–404.

[Elmasri and Navathe, 2011] Elmasri, R. and Navathe, S. B. (2011). *Fundamental of database systems*. Addison-Wesley, sexth edition.

[Eslami and Nazami, 2011] Eslami, S. and Nazami, E. (2011). An automatic approach for topic maps development using relational databases. In *International conference on computer research and development*. IEEE Press.

[Garshol, 2002] Garshol, L. M. (2002). What are topic maps. [Online: http://www.xml.com/pub/a/2002/09/11/topicmaps.html, Visitado en: Julio-2012].

[Garshol, 2006] Garshol, L. M. (2006). tolog – a topic maps query language. In *Charting the Topic Maps Research and Applications Landscape*, pages 183–196. Springer Berlin.

[Gilleson, 2005] Gilleson, M. L. (2005). *Fundamental of database managment systems*. John Wiley and Sons, Inc.

[Grand and Soto, 2002] Grand, B. L. and Soto, M. (2002). Visualisation of the semantic web: Topic maps visualisation. In *Proceedings of the Sixth International Conference on Information Visualisation*. IEEE Press.

[Gronmo, 2000] Gronmo, G. O. (2000). Creating semantically valid topic maps. In *XML Europe*.

[Gronmo, 2002] Gronmo, G. O. (2002). Creating topic maps from existing data sources.

[Gylta and Barta, 2002] Gylta, J. and Barta, R. (2002). Xtmpath, manipulating topic map data structures. [Online: `http://topicmaps.bond.edu.au/docs/13?style=printable`, Visitado en: Julio-2012].

[Habert and Folch, 2002] Habert, B. and Folch, H. (2002). Articulating conceptual spaces using the topic map standard. In *Proceedings XML*, pages 8–13.

[Hai-yun and Shu-feng, 2010] Hai-yun, L. and Shu-feng, Z. (2010). Translating relational databases into rdf. In *International Conference on Environmental Science and Information Application Technology*, volume 3, pages 464–467.

[Heru Agus Santoso, 2011] Heru Agus Santoso, Su-Cheng Haw, Z. T. A.-M. (2011). Ontology extraction from relational database: Concept hierarchy as background knowledge. *Knowledge-Based Systems*, (24):457–464.

[ISO/IEC, 1999] ISO/IEC (1999). Iso/iec 13250, topic maps. [Online: `http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0129.pdf`, Visitado en: Mayo-2012].

[ISO/IEC, 2004] ISO/IEC (2004). Topic maps query language. [Online: `http://www.isotopicmaps.org/tmql/`, Visitado en: Marzo-2012].

[ISO/IEC, 2006a] ISO/IEC (2006a). Topic maps data model. [Online: `http://www.isotopicmaps.org/sam/sam-model/`, Visitado en: Marzo-2012].

[ISO/IEC, 2006b] ISO/IEC (2006b). Topic maps - xml syntax. [Online: http://www.isotopicmaps.org/sam/sam-xtm/. Visitado en: Marzo-2012].

[ISO/IEC, 2007] ISO/IEC (2007). Topic maps - graphical notation. [Online: http://www.isotopicmaps.org/gtm/, Visitado en: Marzo-2012].

[Jung-Mn et al., 2007] Jung-Mn, K., Hyopil, S., and Hyoung-Joo, K. (2007). Schema and constraints-based matching and merging of topic maps. *Information Processing and Management*, (43):930–945.

[Kasler et al., 2006] Kasler, L., Venczel, Z., and Varga, L. Z. (2006). Framework for semi automatically generating topic maps. In *Proceedings of the 3rd international workshop on text-based information retrieval*, pages 24 – 30.

[Laclavík, 2006] Laclavík, M. (2006). Rdb2onto: Relational database data to ontology individuals mapping. In *Tools for Acquisition, Organisation and Presenting of Information and Knowledge*, pages 86–89.

[Lavik and Nordeng, 2007] Lavik, S. and Nordeng, T. W. (2007). Remote topic maps in learning. In *Second International Conference on Topic Maps Research and Applications, TMRA*, pages 67–73. Springer-Verlag.

[Lee et al., 2001] Lee, T. B., Hendler, J., and Lassila, O. (2001). The semantic web. *Scientific American*. [Online: http://campus.fsu.edu/bbcswebdav/users/bstvilia/lis5916metadata/readings/scientific-american_0.pdf, Visitado en: Junio-2012].

[Librelotto et al., 2004] Librelotto, G. R., Ramalho, J. C., and Henriques, P. R. (2004). Tm-builder: An ontology builder based on xml topic maps. *Clei electronic journal*, 7(2).

[Lin and Qin, 2002] Lin, X. and Qin, J. (2002). Building a topic map repository.

[Mark, 2007] Mark, G. (2007). Semantic web 2.0. In *Intelligent Systems, IEEE*, volume 22, pages 94–96.

[Moore and Ahmed, 2005] Moore, G. and Ahmed, K. (2005). Topic map relational query language tmrql. Technical report, Networked Planet Limited.

[Neidhart et al., 2009] Neidhart, T., Pinchuk, R., and Valentin, B. (2009). Semantic integration of relational data sources with topic maps. In *Topic Maps Research and Applications*. Springer-Verlag Berlin Heidelberg.

[Ontopia, 2010] Ontopia (2010). The ontopia vizigator. [Online: `http://www.ontopia.net/doc/5.2.1/vizigator/userguide.html`, Visitado en: Abril-2012].

[Ontopia, 2011] Ontopia (2011). Db2tm user's guide. [Online: `http://www.ontopia.net/doc/5.2.1/db2tm/user-guide.html`, Visitado en: Mayo-2012].

[Pepper, 2002] Pepper, S. (2002). The tao of topic maps. [Online: `http://www.ontopia.net/topicmaps/materials/tao.html`, Visitado en: Mayo-2012].

[Pepper, 2012] Pepper, S. (2012). Ontopia: Open source tools for building, maintaining and deployment topic maps-base applications. [Online: `http://www.ontopia.net`, Visitado en: Mayo-2012].

[Rani et al., 2007] Rani, P., Richard, A., Juan-Jose, D. O., Els, D., David, D. W., Georges, F., and Bernard, F. (2007). Toma: Tmql, tmcl, tmml. In *Proceedings of the 2nd international conference on Topic maps research and applications*, pages 107–129. Springer-Verlag.

[Redmann et al., 2008] Redmann, T., Thomas, H., Markscheffel, B., and Pressler, M. (2008). Gtmalpha towards a graphical notation for topic maps. In *Fourth International Conference on Topic Maps Research and Applications*.

[Reynolds and Kimber, 2000] Reynolds, J. and Kimber, W. E. (2000). Topic map authoring with reusable ontologies and automated knowledge mining. In *XML 2002 Proceedings by deepX*.

[Roberson and Dicheva, 2007] Roberson, S. and Dicheva, D. (2007). Semi-automatic ontology extraction to create draft topic maps. In *Proceedings of the 45th annual southeast regional conference*, pages 100 – 105.

[Wandora, 2012] Wandora (2012). Wandora: the information extraction, management and publishing application. [Online: `http://www.wandora.org/`, Visitado en: Junio-2012].

[Ye et al., 2011] Ye, F., Li, H., and Hu, M. (2011). The construct of data integration model of heterogeneous e-government system based on topic maps. In *International Conference on Intelligent Computation Technology and Automation*. IEEE Press.

CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL IPN

UNIDAD TAMAULIPAS

Cd. Victoria, Tamaulipas, a 29 de octubre de 2012.

Los abajo firmantes, integrantes del jurado para el examen de grado que sustentará el C. ADAN JOSE GARCIA, declaramos que hemos revisado la tesis titulada:

"Obtención de Topic Maps a partir de Bases de Datos Relacionales"

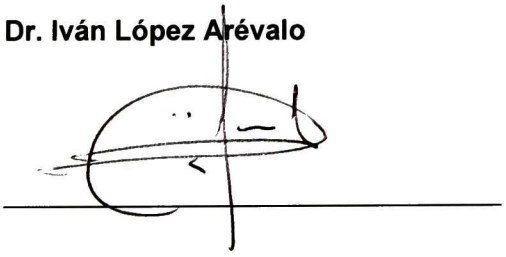
Y consideramos que cumple con los requisitos para obtener el grado de Maestro en Ciencias en Computación.

Atentamente,

Dr. J. Guadalupe Rodríguez García

Dr. Javier Rubio Loyola

Dr. Iván López Arévalo



Km. 5.5 Carretera a Soto la Marina. Ciudad Victoria, Tamaulipas. CP 87130
Parque Científico y Tecnológico TECNOTAM
Tels.: (834) 1070220 y 316 6600
E-mail: admin@tamps.cinvestav.mx