Universidad
Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

Linked Data para la Generación de Conocimiento Financiero a partir de la Extracción de Información Semiestructurada

Autor:

José Luis Sánchez Cervantes

Director/es:

Dr. Juan Miguel Gómez Berbís

Dr. José Luis López Cuadrado

Tutor:

Dr. Juan Miguel Gómez Berbís

DEPARTAMENTO DE INFORMÁTICA

Leganés, Noviembre 2014

Universidad Carlos III de Madrid
www.uc3m.es

TESIS DOCTORAL

LINKED DATA PARA LA GENERACIÓN DE CONOCIMIENTO FINANCIERO A PARTIR DE LA EXTRACCIÓN DE INFORMACIÓN SEMIESTRUCTURADA

Autor: José Luis Sánchez Cervantes

Director/es: Dr. Juan Miguel Gómez Berbís

Dr. José Luis López Cuadrado

Firma del Tribunal Calificador:

Firma

Presidente:

Vocal:

Secretario:

Calificación:

Leganés/Getafe, de de

Summary

At present, the information is generated by data located in a distributed environment but linked. In relation to this premise, semantic and Linked Data technologies provide a paradigm in which not only documents but also the data are first-class resources on the Web, allowing its extension and sharing of knowledge towards a global space data based on open standards, better known as Web of data.

In this thesis, a semantic model based on the principles of Linked Data that provides an alternative solution to the problems presented data integration that are manifested in the financial statements published by the company under the XBRL standard through Web. In this sense, using the semantic model are identified and remedied certain limitations in Balance Sheets, Income Statements and Cash Flow Statements. These limitations include the lack of a semantics that allow the integration of their data to make it navigable, difficulty accessing them through associated Internet protocols like HTTP for navigation and interconnection with other sources of information and lacking ability for search of financial ratios, as well as processing calculations that allow a fundamental or classical financial analysis that supports the decision making.

The semantic model integrated financial taxonomies based on US-GAAP standard, and is the main support base of reusable financial knowledge inspired Linked Data. In connection with this, the research conducted in this thesis is synthesized through the solution provided to the hypothesis raised therein, and by which seeks to demonstrate that the semantic model has the ability to financial populate a knowledge base from the integration of external data sources, facilitating the reuse of data with third parties via Linked data, help improve the structural quality of financial data and verify that this model also facilitates analysis financial crucial to support decision-making both automated and by the people.

Resumen

En la actualidad, la información es generada por datos ubicados en un entorno distribuido pero vinculado. Con relación a esta premisa, las tecnologías semánticas y Linked Data, proporcionan un paradigma en el que no sólo los documentos, sino que también los datos son recursos de primera clase en la Web, permitiendo su extensión y la compartición de conocimientos hacia un espacio global de datos basado en estándares abiertos, mejor conocido como Web de datos.

En este trabajo de tesis, se presenta un modelo semántico inspirado en los principios de Linked Data que ofrece una alternativa de solución a los problemas de integración de datos que se manifiestan en los Estados financieros publicados por las empresas bajo el estándar XBRL a través de la Web. En este sentido, mediante el modelo semántico se identifican y subsanan ciertas limitaciones existentes en las Hojas de Balance, Cuentas de resultados y Estados de flujos de efectivo. Entre estas limitaciones destacan la falta de una semántica que permita la integración de sus datos para hacerlos navegables, la dificultad para el acceso a los mismos a través de protocolos asociados a Internet como el HTTP para la navegación e interconexión con otras fuentes de información, y la carente capacidad para la búsqueda de ratios financieros, así como el procesamiento de cálculos permitan un análisis fundamental o clásico que sirva de apoyo a la toma de decisiones.

El modelo semántico, se integra de taxonomías financieras basadas en la norma US-GAAP, y es el soporte fundamental de una base de conocimientos financieros reutilizable inspirada en Linked Data. En relación con lo anterior, la investigación que se realiza en este trabajo de tesis se sintetiza a través de la solución que se proporciona a las hipótesis que en ella se plantean, y mediante la que se busca demostrar que el modelo semántico tiene la capacidad para poblar una base de conocimientos financieros a partir de la integración de fuentes de datos externas, facilitar la reutilización de sus datos con terceros a través de Linked Data, ayudar a mejorar la calidad estructural de los datos financieros y comprobar que este modelo también facilita el análisis fundamental financiero para apoyar la toma de decisiones tanto automatizada como por parte de las personas.

Índice general

1. Introducción 1

1.1 Motivación de la investigación 3

1.2 Justificación 5

1.2.1 ¿Cúal es la importancia de estudiar la información financiera publicada en la Web para la generación de conocimiento? 5

1.2.2 ¿Por qué centrarse en el paradigma Linked Data? 7

1.2.3 ¿Por qué se plantea un modelo semantico de datos financieros inspirado en Linked Data? 8

1.2.4 ¿Qué aportaciones tendrá esta tesis doctoral? 9

1.3 Objetivo general y objetivos específicos 10

1.3.1 Objetivo General 10

1.3.2 Objetivos Específicos 10

1.4 Organización de la tesis doctoral 11

2. Estado del arte 13

2.1 Análisis fundamental financiero 13

2.1.1 Ratios financieros 13

2.1.2 Estados financieros 19

2.2 Normas internacionales de información financiera 20

2.2.1 Normas internacionales US-GAAP e IFRS 21

2.3 El lenguaje XBRL 22

2.3.1 XML como lenguaje universal y abierto 23

2.3.2 XBRL como lenguaje estándar para la presentación de información financiera 25

2.3.2.1 Expresión de reglas con XBRL 25

2.3.3 Estructura básica de los Estados financieros basados en XBRL 26

2.3.4 Transparencia financiera con XBRL 28
2.4 Estado del arte de las tecnologías semánticas 30
2.4.1 Web Semántica 30
2.4.2 Ontologías 32
2.4.3 Lenguajes para el desarrollo de ontologías 33
2.4.4 Herramientas para el desarrollo de aplicaciones de la Web Semántica 43
2.4.5 Modelo Conceptual 47
2.4.6 Modelos de datos 48
2.4.6.1 Modelo de datos entidad-relación 48
2.4.6.2 Modelo relacional 49
2.4.6.3 Otros modelos de datos 49
2.4.7 Modelos de datos Semánticos 50
2.4.7.1 Modelo de datos Entidad-Atributo-Valor 52
2.4.7.2 Modelo de datos común 53
2.4.8 Linked Data 54
2.4.8.1 La importancia de Linked Data 55
2.4.8.2 Publicación de Linked Open Data en la Web 56
2.4.9 Linked Open Data cloud 57
2.4.9.1 DBpedia como núcleo de la Linked Open Data cloud 59
2.4.10 Repositorios RDF 60
2.4.11 Herramientas para la publicación de Linked Data 63
2.4.12 Herramientas para el descubrimiento de enlaces RDF en la Web de datos 65
2.4.13 Lenguajes para la consulta de grafos RDF 67
2.4.14 Ciclos de vida de Linked Data 71
2.5 Trabajos relacionados 76
2.5.1 Extracción, procesamiento y representación de datos financieros 76
2.5.2 Trabajos para el análisis de la información financiera 79
2.6 Conclusiones del estado del arte 85

3. Planteamiento de las hipótesis a resolver 88

3.1 Propuesta de validación de las hipótesis 89

3.2 Proceso de transformación de datos financieros para la validación de las hipótesis planteadas 90

4. Modelado semántico de datos 95

4.1 Modelo Mixto de datos financieros 95

4.2 Taxonomías de los modelos semánticos de datos financieros 100

4.3 Interconexión del modelo Mixto de datos financieros 112

4.4 Trancisión de la conceptualización del modelo semántico hacia la base de conocimientos financieros inspirada en Linked Data 115

4.5 Modelo Entidad-Atributo-Valor de datos financieros inspirado en Linked Data 119

4.6 Navegabilidad de los modelos semánticos 120

5. Base de conocimientos financieros inspirada en Linked Data 123

5.1 Estadísticas de los Estados financieros XBRL descargados para su transformación en RDF 123

5.2 Infraestuctura tecnológica utilizada para el almacenamiento de las tripletas RDF transformadas 124

5.3 Estadísticas de las tripletas RDF almacenadas en la base de conocimientos financieros 126

6. Introducción y validación de los modelos semánticos de datos financieros inspirados en Linked Data 129

6.1 Análisis comparativo entre los modelos semánticos Mixto y Entidad-Atributo-Valor 130

6.1.1 Diseño de consultas basadas en SPARQL para el análisis comparativo entre los modelos semánticos 130

6.1.2 Resultados del análisis comparativo entre los modelos semánticos 132

6.1.3 Conclusión del análisis comparativo entre los modelos semánticos 134

6.2 Descubrimiento y vinculación de datos financieros en la Web 135

6.2.1 Elección del marco de trabajo para la ejecución de los experimentos para el descubrimiento de enlaces en la LOD cloud 136

6.2.2 Experimentos para el descubrimiento de enlaces con información financiera en la LOD cloud 137
6.2.3 Validación de los enlaces con información financiera descubiertos en la LOD cloud 139
6.2.4 Conclusión del descubrimiento y vinculación de datos financieros en la Web 142
6.3 Caso de estudio: análisis comparativo de empresas por sector para la recomendación de una inversión 142
6.3.1 Razón corriente de las empresas Wal-Mart y Costco 144
6.3.1.1 Razón corriente de la empresa Wal-Mart 144
6.3.1.2 Razón corriente de la empresa Costco 145
6.3.2 Capital de trabajo de las empresas Wal-Mart y Costco 147
6.3.3 Prueba ácida de las empresas Wal-Mart y Costco 149
6.3.4 Razón de deuda para las empresas Wal-Mart y Costco 151
6.3.4.1 Razón de deuda para la empresa Wal-Mart 152
6.3.4.2 Razón de deuda para la empresa Costco 153
6.3.5 Obtención de información adicional de las empresas 155
6.3.6 Conclusiones del Caso de estudio 155
6.4 Validación de resultados a través de expertos 158
6.4.1 Resultados y discusión 160
6.5 Conclusión del proceso de validación 163
7. Conclusiones y futuras líneas de investigación 166
7.1 Conclusiones 166
7.2 Líneas futuras de investigación 168
Anexos 170
Publicaciones realizadas 170
Acrónimos 171
Bibliografía 174

Índice de figuras

Figura 1. Estructura de las taxonomías que integran a los Estados financieros publicados en XBRL 26

Figura 2. Tecnologías Web y tecnologías de la Web Semántica 31

Figura 3. Elementos de una tripleta basada en RDF 34

Figura 4. Pila de lenguajes para el desarrollo de ontologías.. ..... .42

Figura 5. Diagrama de la Linked Open Data cloud (LOD cloud) hasta Abril de 2014 58

Figura 6. Ciclo de vida de Linked Data de acuerdo a Hyland et al., (2011) 71

Figura 7. Ciclo de vida de Linked Data de acuerdo a M Hausenblas et al., (2011) 71

Figura 8. Ciclo de vida de Linked Data iterativo de acuerdo a Villazón-Terrazas et al., (2011) 71

Figura 9. Ciclo de vida de Linked Data para el proyecto LOD 2 73

Figura 10. Proceso de publicación de Linked Data visto como un ascensor de datos 75

Figura 11. Proceso para la generación de la base de conocimientos financieros inspirada en Linked Data 89

Figura 12. Modelo semántico Mixto de datos financieros inspirado en Linked Data 96

Figura 13. Taxonomía RDF(S) de Hoja de balance *(Balance sheet)* basada en la norma US-GAAP 101

Figura 13.1 Ratios (Subclases) del Activo circulante *(Current Assets)* de la taxonomía de Hoja de balance 102

Figura 13.2 Ratios (Subclases) del Activo fijo *(Current Assets)* de la taxonomía de Hoja de balance 103

Figura 13.3 Ratios (Subclases) de los Pasivos *(Liabilities)* de la taxonomía de Hoja de balance 104

Figura 13.4 Ratios (Subclases) del Capital de socios incluyendo la parte atribuible a la participación no controladora *(Partners' Capital, Including Portion Attributable to Noncontrolling Interest)* de la taxonomía de Hoja de balance 104

Figura 13.5 Ratios (Subclases) del Capital de socios, número de unidades, valor nominal y otras cláusulas *(Partners' Capital, Number of Units, Par Value and Other Disclosures)* de la taxonomía de Hoja de balance 104

Figura 13.6 Ratios (Subclases) de la Equidad temporal *(Temporary Equity)* de la taxonomía de Hoja de balance 105

Figura 13.7 Ratios (Subclases) del Capital contable, número de acciones, valor nominal y otras cláusulas *(Stockholders' Equity, Number of Shares, Par Value and Other Disclosures)* de la taxonomía de Hoja de balance 105

Figura 13.8 Ratios (Subclases) del Capital contable, incluyendo la porción atribuible a la participación no controladora *(Stockholders' Equity, Including Portion Attributable to Noncontrolling Interest)* de la taxonomía de Hoja de balance 105

Figura 14. Taxonomía RDF(S) de Estado de flujo de efectivo *(Cash flow)* basada en la norma US-GAAP 106

Figura 15. Taxonomía RDF(S) de la Cuenta de estado de resultados *(Income statement)* basada en la norma US-GAA 106

Figura 16. Modelo semántico Mixto de datos financieros inspirado en Linked Data e interconexión de datos 112

Figura 17. Ejemplo de presentación de los Estados financieros de APPLE INC. 116

Figura 18. DEI de APPLE INC. Para el periodo del 29 Marzo de 2014 118

Figura 19. Conceptualización del modelo EAV de datos financieros con reificación 119

Figura 20. Representación de navegabilidad en el modelo EAV 120

Figura 21. Representación de navegabilidad en el modelo semántico Mixto 121

Figura 22. Porcentajes correspondientes a las tripletas RDF de la base de conocimientos financieros 126

Figura 23. Gráfica trimestral de la Razón corriente para Wal-Mart 144

Figura 24. Datos trimestrales de la Razón corriente para Wal-Mart 145

Figura 25. Gráfica trimestral de la Razón corriente para Costco 146

Figura 26. Datos trimestrales de la Razón corriente para Costco 146

Figura 27. Datos trimestrales del Capital de trabajo para Wal-Mart 147

Figura 28. Datos trimestrales del Capital de trabajo para Costco 148

Figura 29. Datos trimestrales de la Prueba ácida para Wal-Mart 150

Figura 30. Datos trimestrales de la Prueba ácida para Costco 150

Figura 31. Pasivos Corrientes de Wal-Mart..... 152

Figura 32. Deuda a largo plazo, vencimientos actuales totales de Wal-Mart 152

Figura 33. Activos Corrientes de Wal-Mart 153

Figura 34. Depreciación acumulada, agotamiento y amortización de propiedades, planta y equipo de Wal-Mart 153

Figura 35. Pasivos Corrientes de Costco 154

Figura 36. Deuda a largo plazo, vencimientos actuales totales de Costco..... 154

Figura 37. Activos Corrientes de Costco..... 154

Figura 38. Depreciación acumulada, agotamiento y amortización de propiedades, planta y equipo de Costco..... 154

Figura 39. Enlaces a DPpedia a través del visualizador del conjunto de datos financieros..... 155

Índice de tablas

Tabla 1. Ratios de rotación 16

Tabla 2. Ratios de liquidez 17

Tabla 3. Ratios de Solvencia 18

Tabla 4. Ejemplo de información ordinaria y metadatos 24

Tabla 5. Ratios de la taxonomía de Hoja de balance 108

Tabla 5 (Continuación). Ratios de la taxonomía de Hoja de balance 109

Tabla 5 (Continuación). Ratios de la taxonomía de Hoja de balance 110

Tabla 5 (Continuación). Ratios de la taxonomía de Hoja de balance 111

Tabla 6. Detalle de los Estados financieros descargados hasta el tercer trimestre de 2014 124

Tabla 7. Infraestructura tecnológica utilizada para el repositorio de datos financieros 125

Tabla 8. Consultas basadas en SPARQL para el *Benchmarking* de los modelos semánticos 131

Tabla 9. Características de las consultas basadas en SPARQL utilizadas en el *benchmarking* 132

Tabla 10. Tiempos de adquisición de datos para los modelo semánticos Mixto y EAV 133

Tabla 11. Resultados del descubrimiento de enlaces en la LOD cloud a través de DBpedia 138

Tabla 12. Resultados de la validación de los enlaces descubiertos 141

Tabla 13. Razón corriente por año de la empresa Wal-Mart 145

Tabla 14. Razón corriente por año de la empresa Costco 146

Tabla 15. Capital de trabajo por año de la empresa Wal-Mart 148

Tabla 16. Capital de trabajo por año de la empresa Costco 149

Tabla 17. Razón de deuda por año de la empresa Wal-Mart 153

Tabla 18. Razón de deuda por año de la empresa Costco 154

Tabla 19. Respuestas de expertos a preguntas del cuestionario 160

Tabla 20. Opiniones y recomendaciones proporcionadas por los expertos 162

Capítulo 1

Introducción, motivación y objetivos

Resumen. La ciencia y tecnología informática se vincula con innumerables áreas del conocimiento, siendo las finanzas, la toma de decisiones y el cómo estas se benefician a través de las tecnologías semánticas y el paradigma Linked Data, el punto de partida para la investigación con el que se da inicio a la presente tesis doctoral. El primer capítulo, inicia proporcionando una introducción acerca de esta investigación, seguido de los motivos que rigen a su realización, y continuando con la descripción de los objetivos general y específicos que llevarán a su correcta culminación. Finalmente, el presente capítulo proporciona una breve descripción acerca de la estructura de la tesis doctoral.

1. Introducción

Con el transcurso de los años la informática ha sido vinculada con numerosas áreas del conocimiento entre las que se incluyen las finanzas y la contabilidad, siendo éstas y su relación con Linked Data, el origen de la investigación que se detalla en la presente tesis doctoral. En los últimos años, el ecosistema de datos financieros se está beneficiando de la creciente cantidad de datos publicados en la Web a través de documentos divulgados por múltiples organizaciones empresariales, generando importantes fuentes de información. Esto ocurre gracias a que existen prácticas contables regidas por normas que obligan a las empresas a divulgar públicamente su información financiera, un ejemplo de ello es la norma *Sarbanes-Oxley*[1] de la U.S. SEC[2] (*USA-Securities and Exchange Commission*, Comisión del Mercado de Valores de los Estados Unidos de América) en la que se indican una serie de reformas para mejorar la responsabilidad de las empresas, mejorar la divulgación de la información financiera y combatir el fraude corporativo y de contabilidad. Por lo tanto, las empresas requieren de soluciones que permitan el análisis de datos de carácter financiero con la finalidad de adquirir todos los beneficios posibles. En este sentido, la importancia del intercambio de datos financieros radica en fomentar su reutilización generando alternativas para el desarrollo, mejora y uso de herramientas informáticas que permitan emitir y recibir información financiera, consolidarla y analizarla para facilitar el apoyo a la toma de decisiones tanto automatizada como por parte de las personas. Con base en esto, se han realizado importantes esfuerzos para que el intercambio de datos financieros se realice de una manera más estructurada y computable. En particular la iniciativa XBRL (*eXtensible Business Reporting Language,* Lenguaje Extensible de Informes de Negocios)[3] (Hoffman & Van-Egmond, 2012), impulsó un enfoque estandarizado para la publicación de Estados financieros por parte de las empresas que utilizan tecnologías basadas en XML (*Extensible Markup Language,* Lenguaje de Marcas Extensible) (Bray et al., 1998) y que están obligadas a publicar sus Estados financieros, entre los que sobresalen las Hojas balance, el Estado de flujo de efectivo y la Cuenta de estado de resultados. En el ámbito de la contabilidad financiera, las Hojas de balance presentan un resumen de la situación financiera de una empresa unipersonal (en este contexto, una empresa unipersonal es aquella que se conforma de una sola persona), una sociedad, una corporación u otra organización empresarial en el que los activos, pasivos y patrimonio de propiedad se muestran a partir de una fecha específica. En resumen, un Balance general es descrito

[1]Sarbanes-Oxley: http://www.sec.gov/about/laws.shtml#sox2002
[2]U.S. SEC: http://www.sec.gov/
[3]XBRL: http://www.xbrl.org/

como un "*informe instantáneo de la situación financiera de una empresa"* (Williams et al., 2005). La iniciativa XBRL tiene un importante papel en la divulgación de los Estados financieros y abre nuevas posibilidades para el intercambio de datos entre aplicaciones informáticas (Hoffman & Van-Egmond, 2012). Aunque proporcionar esquemas para la representación estructurada de datos financieros resulta en una mejora significativa, no resuelve los problemas de integración de esos datos por la ausencia de una semántica que les permita mantenerse interrelacionados e integrados (H. Zhu & Madnick, 2007). Es decir, que los datos no son navegables entre ellos y no permiten su interconexión con fuentes de datos externas, teniendo como resultado ciertas limitaciones en el desarrollo de tareas financieras esenciales y de interés para las empresas, tales como, la generación de una base conocimientos financieros que proporcione capacidades para la interconexión y navegabilidad entre sus datos y que sea interoperable con otras fuentes de información relacionada. Además, sería necesario que permita el análisis y el cálculo de indicadores financieros adicionales que sirvan de apoyo para la toma de decisiones tanto automatizada como por parte de las personas (Bartley, Chen, & Taylor, 2011; T. Harris & Morsfield, 2012). En este contexto, la interoperabilidad es la habilidad que tienen dos o más sistemas o componentes para intercambiar información, y hacer uso de la información intercambiada (IEEE, 1990).

Este trabajo de tesis doctoral propone un modelo semántico que se inspira en los principios de Linked Data (T Berners-Lee, 2009) y el uso de tecnologías semánticas para la representación estructurada de datos financieros con el objetivo de interrelacionarlos e integrarlos para hacerlos interoperables con fuentes de datos externas aprovechando los beneficios que la Web ofrece. En este sentido, Larrán-Jorge & Giner, (2002) concluyen que la Web sirve de medio para la divulgación de información empresarial y permite ser utilizada de forma interactiva proporcionando a los usuarios el beneficio de obtener la información que realmente necesitan para tomar sus decisiones. Además, la Web hace posible divulgar la información entre un mayor número de usuarios y de manera más oportuna. Suministrar interoperabilidad a los datos a través de Linked Data, ayuda a superar las deficiencias de los informes publicados bajo el estándar XBRL. Tales deficiencias están inmersas en los propios documentos XBRL, que conservan un formato para la representación estructurada de la información financiera (Plumlee & Plumlee, 2008; Shin, 2003), pero no mantienen una semántica que les permita una integración entre sus datos para ser navegables, no manejan protocolos de acceso asociados a Internet como el Protocolo de Transferencia de Hipertexto *(HTTP, Hypertext Transfer Protocol)* para navegar e

interconectarse con fuentes de datos externas y no posibilitan la capacidad de realizar procesos de inferencia con sus datos. Por otra parte, los Estados financieros que se articulan en ficheros de texto plano, presentan problemas asociados a la gestión de ficheros como la falta de concurrencia, comprobación de integridad y seguridad (XBRL-España, 2006). Por lo tanto, el modelo semántico que se propone, aborda cada una de las deficiencias previamente descritas, incluyendo los problemas de reutilización de datos.

El modelo semántico basado en Linked Data propuesto en esta tesis, involucra esfuerzos en la búsqueda, obtención y categorización de la información financiera publicada por las empresas a través de sus Estados financieros, que a pesar de venir con contenido estructurado o semiestructurado, carecen de una organización lógica para ser tratados como datos. Parte esencial del modelo semántico, se centra en brindar la transformación de los datos contenidos en esos Estados financieros hacia una notación semántica, lo que proporciona la generación de conocimiento financiero, conocimiento que podrá ser aprovechado por otros sistemas económicos y financieros para acceder a fuentes de datos específicas y navegar entre los datos a través de sus relaciones, siguiendo el paradigma de Linked Data, ya que Linked Data proporciona el soporte adecuado para aprovechar las conexiones existentes entre la información financiera.

Una vez tocado el punto de los Estados financieros publicados en la Web con contenido semiestructurado, como es el caso de los documentos basados en el estándar XBRL. Es importante resaltar que durante el transcurso de este trabajo de tesis, los datos publicados bajo este estándar al ser un subconjunto de instrucciones de XML, serán tratados como semiestructurados, aseverando la afirmación hecha por (Broekstra, Klein, et al., 2002; Decker, 2002) en la que mencionan que XML no tiene la expresividad suficiente para la realización de inferencias lógicas.

1.1 Motivación de la investigación

Las empresas que están obligadas a publicar su información financiera y que difunden sus Estados financieros (Hojas balance, Estado de flujo de efectivo y Cuenta de Estado de resultados) en la Web siguiendo el estándar XBRL, generan importantes fuentes de información. A pesar de la existencia de diversos formatos para la divulgación de este tipo de información, tales como ficheros en texto plano, hojas de cálculo y documentos PDF, por mencionar algunos, la adopción de este estándar por parte de las empresas se considera imperativo ya que es un método de bajo costo que mejora su productividad y competitividad, permitiendo la eliminación de la brecha digital existente entre distintos

colectivos y regiones (Pinsker & Li, 2008). Adicionalmente, supone una mayor confianza en los datos que las empresas publican y por lo tanto un incremento en la transparencia empresarial y, consecuentemente, en el acceso a la financiación (XBRL-España, 2005). El hecho de publicar Estados financieros bajo el estándar XBRL genera varias expectativas siendo entre las principales proporcionar una alternativa de solución a los problemas de integración de los datos que se publican mediante dicho estándar. A pesar de esto, este proceso de publicación representa una sobrecarga significativa para las empresas (Sinnett, 2011) ya que no todas respetan el formato de publicación XBRL de manera estricta, lo que proporciona una solución parcial a estos problemas, es decir, que el contenido de los Estados financieros basados en XBRL se encuentra semiestructurado y requiere de un proceso de transformación semántica para ofrecer una presentación de datos fácil de leer e interpretar (Plumlee & Plumlee, 2008; Shin, 2003). Adicionalmente, XBRL carece de semántica y por lo tanto los datos publicados bajo este estándar no mantienen una integración que les permita ser navegables e interoperables con fuentes de datos externas (H. Zhu & Madnick, 2007), restringiendo las posibilidades para la realización del análisis e investigación holística de otras fuentes tangibles de información financiera (O'Riain, Curry, & Harth, 2012). Estos datos carecen de una pre-interpretación que les proporcione orden y sentido, lo que significa que los datos por sí mismos no tienen la capacidad de comunicar algún significado y no son idóneos para ser computados en un entorno automático con capacidades para el razonamiento de datos financieros, el cual es fundamental para la toma de decisiones rápidas que estén basadas en información precisa y fiable (O'Riain, Harth, & Curry, 2012). Por lo tanto, la publicación de datos basados en el estándar XBRL seguirá jugando un papel muy importante, papel que resulta conveniente ya que a través de la semántica, los datos publicados en ese estándar son mejorados con características de expresividad, integridad e interoperabilidad (Grosof, 2009).

La principal motivación para realizar esta tesis doctoral, consiste en razonar y dar solución a las limitaciones mencionadas en el párrafo anterior. Si consideramos las tendencias actuales para la mejora del estándar XBRL (T. Harris & Morsfield, 2012), el modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009) que se plantea, no se orienta directamente a la mejora de los Estados financieros publicados bajo tal estándar, sino que propone una manera de impulsar la integración e interoperabilidad de los datos financieros a través de taxonomías financieras que corresponden a las Hojas de Balance, Estado de flujo de efectivo, Estado de cuenta de resultados y formatos altamente escalables, sirviendo de base para ofrecer una arquitectura flexible y viable para la

automatización de tareas financieras tales como el análisis de datos financieros, el cálculo de ratios financieros adicionales, servir de apoyo en la toma de decisiones tanto automatizada como por parte de las personas, apoyar en la mitigación de riesgos en la reputación de las empresas, gestionar la información financiera, generar razonamiento automático en los datos financieros, servir de fuente de información para otros sistemas financieros, y facilitar la realización de descubrimientos financieros propios, entre otras tareas.

1.2 Justificación

Como se ha indicado en la sección anterior, la publicación de los Estados financieros bajo el estándar XBRL genera importantes fuentes de información por parte de las empresas y abre nuevas posibilidades para el intercambio de datos entre aplicaciones informáticas (Hoffman & Van-Egmond, 2012). Sin embargo, a pesar de presentar los datos financieros de manera estructurada, los Estados financieros XBRL requieren de un proceso de transformación hacia un formato legible utilizando, por ejemplo, Transformaciones XSLT *(eXtensible Stylesheet Language Transformations)*, que es un lenguaje para la transformación de documentos XML (*eXtensible Markup Language,* Lenguaje de Marcas Extensible), o a través del uso de Hojas de Estilo en Cascada *(CSS, Cascading Style sheets)*. Tanto XSLT como CSS, son estándares XML y sólo se utilizan para la presentación de los datos (Richards, Smith, & Saeedi, 2006). Esto significa que a nivel de datos, los Estados financieros XBRL carecen de semántica limitando potencialmente su aprovechamiento para integrarlos e interconectarlos para hacerlos interoperables con fuentes de información externas (H. Zhu & Madnick, 2007). En este sentido, a continuación se plantean las cuestiones que fundamentan la presente tesis doctoral.

1.2.1 ¿Cúal es la importancia de estudiar la información financiera publicada en la Web para la generación de conocimiento?

La importancia del estudio relacionado con la información financiera publicada en la Web se fundamenta en los avances que se tienen desde que las organizaciones empresariales utilizan a la Web como medio para la divulgación de su información financiera, para que esta sea aprovechada convenientemente por un mayor número de usuarios (Larrán-Jorge & Giner, 2002). De acuerdo con lo expresado, en este punto radica el éxito de la Web, ya que entre otras cosas, facilita el acceso a una enorme cantidad de contenidos en constante crecimiento y tiene la capacidad de establecer comunicaciones de diversos tipos a bajo coste (Jiménez-Domingo, 2013). Por otra parte, la mayoría de las

representaciones digitales de información financiera publicadas en la Web se han codificado en HTML (*HyperText Markup Language,* Lenguaje de Marcado de Hipertexto), que controla la forma en que se muestra la información, en términos de apariencia, tamaño, forma, y color (Shin, 2003). Sin embargo, por su nula semántica, HTML no reconoce el contenido, por lo que su uso generalmente es limitado y no es eficaz para la extracción de datos. Además, HTML no permite la búsqueda, el análisis o la manipulación de información sin tener que introducir o añadir los datos a partir de una hoja de cálculo o mediante la descarga de alguna otra aplicación de Software con capacidades de análisis y manipulación de datos (Richards et al., 2006; Shin, 2003).

Dadas las deficiencias del HTML para la búsqueda, el análisis y manipulación de datos financieros, en 1998 se utilizó XML como alternativa de solución para solventar tales deficiencias. XML sin la incorporación de esquemas, proporciona capacidades para la representación de información financiera de forma semiestructurada (Broekstra, Klein, et al., 2002; Decker, 2002) y jerárquica, con la identificación única de atributos financieros en relación con una variedad de características de identificación, incluyendo transacciones atómicas y elementos de los Estados financieros como ratios, entidades, períodos de publicación, tipos de unidad (moneda) y los GAAP (*Generally Accepted Accounting Principles*, Principios de Contabilidad Generalmente Aceptados), entre otros (Debreceny & Gray, 2001). A pesar del uso de XML para la presentación estructurada de datos financieros, en un trabajo de investigación presentado por Debreceny, Gray, & Barry (1998), se concluyó la necesidad de desarrollar un conjunto estándar de especificaciones exclusivas para la presentación de Estados financieros y de negocios basados en la Web mediante el uso de XML, así mismo, el AICPA (*American Institute of Certified Public Accountants,* Instituto Americano de Contadores Públicos Certificados) llegó la misma conclusión y proporcionó el capital inicial para investigar y desarrollar una especificación basada en XML para la publicación de información financiera en la Web, teniendo el primer prototipo a finales de Diciembre de 1998. De esta manera surge el estándar XBRL, como un subconjunto de XML (Richards et al., 2006) y como XBRL una de las muchas técnicas de sintaxis para la publicación de informes financieros digitales (Hoffman & Van-Egmond, 2012).

La diferencia sustancial entre un documento de instancia XBRL y un documento XML "puro", radica en que XBRL ofrece un método para expresar operaciones aritméticas de tipo *"Assets = Liabilities + Equity"*, dicho de otra manera, es un método que permite expresar operaciones de contabilidad (Arndt et al., 2006). Además, permite el modelado de

conexiones semánticas (Binstock et al., 2005). De acuerdo con la arquitectura de referencia del estándar XBRL, una taxonomía consiste en un XSD *(XML Schema Definition)* en el que se estructuran y registran los nombres de los *XMLelement*, y los *linkbases* de las tareas correspondientes (etiqueta, cálculo, definición, referencias y presentación, entre otras). En este sentido, el modelado semántico de XBRL se centra en las conexiones entre los *XMLelements* y los *linkbases* a través de *XLinks* (Arndt et al., 2006). Es importante subrayar que este tipo de semántica no es la misma que se aporta con el presente trabajo de tesis, ya que esta iniciativa se centra en proporcionar a los datos financieros una semántica que les permita mantenerse integrados para ser calculables, navegados e interconectados con fuentes de información externa relacionada con ellos, de modo que se potencialice su uso.

Como se ha descrito a lo largo de esta sección, los avances de HTML hacia XML y XBRL para la publicación de la información financiera, se han centrado en la manera en la que son presentados los datos financieros a los usuarios, sin aportar funcionalidades adicionales a la estructuración de los datos financieros mediante esquemas y taxonomías con una semántica débil, como es el caso de XBRL (Arndt et al., 2006; Binstock et al., 2005). En este aspecto y como ya se ha mencionado, la débil semántica de XBRL no permite una integración entre sus datos, que posibilite la navegación y el cálculo directo sobre estos, así como la interconexión con fuentes de datos externas que favorezcan a la interoperabilidad (H. Zhu & Madnick, 2007). Solventar estas deficiencias es lo que se aborda en este trabajo de tesis, de manera que a través de un modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009), se integren los datos financieros de diversas fuentes de información en una base de conocimientos reutilizable que puede ser aprovechada por las personas y empresas para la realización de análisis e investigación de todo lo relacionado con los datos financieros integrados.

1.2.2 ¿Por qué centrarse en el paradigma Linked Data?

El uso de la Web para publicar datos abiertos ha hecho que la información sea más accesible, pero variando los formatos de información se puede dificultar su consumo. Linked Data está basado en el estándar RDF (*Resource Description Framework,* Marco de Descripción de Recursos), que es apto para la presentación de múltiples formatos proporcionando un formato común e interoperable y un modelo para la vinculación e intercambio de datos en la Web (O'Riain, Curry, et al., 2012).

Con Linked Data, la Web se comporta como un ingente espacio de datos distribuidos a nivel mundial (Hartig, Bizer, & Freytag, 2009), y no como una red de documentos, de manera que si los usuarios realizan búsquedas de información obtendrán resultados más precisos. A diferencia de Linked Data, en la Web tradicional los usuarios no realizan búsquedas sobre los datos, los buscadores como *Google Search*, exploran todos los documentos de la red para encontrar las palabras clave de la búsqueda (Brin & Page, 1998), es por eso que muchas de las búsquedas devuelven resultados redundantes. Además, con Linked Data es fácil combinar información de diferentes fuentes sin tener que realizar consultas complejas, funcionalidad que resulta difícil de hacer en la Web tradicional. Son varios los beneficios que se obtienen por la puesta en práctica de Linked Data, siendo algunos de ellos los que se enumeran a continuación (Auer et al., 2013; Heath & Bizer, 2011; W3C, 2011a):

1. Al buscar información directamente sobre datos con identificadores únicos (HTTP URIs), los resultados que se obtienen son más precisos evitando su redundancia y permitiendo la navegación entre los datos de la misma fuente y sobre datos de fuentes de información externas con el fin de obtener información coherente y asociada con el tema de interés del usuario.
2. Linked Data permite contribuciones externas para integrar y combinar datos con otras fuentes de información creadas por usuarios e importantes organizaciones publicas o privadas. En este sentido, los usuarios pueden desarrollar sus propias aplicaciones de Software para consumir y navegar entre los datos de esas fuentes.
3. A través de Linked Open Data, se promueve la transparencia de la información publicada por instituciones públicas y privadas que estén obligadas a divulgar explícitamente sus datos.

Finalmente, Linked Data obtiene mejores resultados si la información se presenta en una forma estándar para fomentar la reutilización. En pocos años, Linked Data revolucionará el mundo del acceso a datos, tomando mucho más importancia en el contexto de la reutilización de datos públicos (Colomo-Palacios et al., 2012).

1.2.3 ¿Por qué se plantea un modelo semantico de datos financieros inspirado en Linked Data?

Según Gonzalez & Dankel, (1993), seguir una serie de pasos correspondientes a una determinada metodología, ayuda a dar solución a un problema. Por otro lado, los modelos

semánticos de datos ayudan a la definición de modelos conceptuales más expresivos mediante la especificación de las relaciones, la abstracción de datos, la herencia, las restricciones, los objetos no estructurados y las propiedades dinámicas de una aplicación (Peckham & Maryanski, 1988). Con base en estas definiciones, se centra la importancia de plantear el modelo semántico de datos financieros que se presenta en esta tesis doctoral, el cual, inspirándose en los principios de Linked Data (T Berners-Lee, 2009), sirve de núcleo para la extracción y transformación en RDF de los datos publicados en documentos XBRL adquiridos a partir de distintas fuentes de información, capturando todos los aspectos de la información financiera expresada en estos documentos y proporcionando capacidades para la realización de tareas analíticas con Estados financieros cruzados así como el procesamiento de consultas versátiles basadas en SPARQL.

1.2.4 ¿Qué aportaciones tendrá esta tesis doctoral?

La investigación que se presenta en esta tesis permitirá la creación de un modelo semántico que se inspira en los principios de Linked Data (T Berners-Lee, 2009) y el uso de tecnologías semánticas para la representación estructurada de datos financieros extraídos a partir de los Estados financieros basados en el estándar XBRL publicados por las empresas, con el propósito de dotar a dichos datos de una semántica que les permita mantenerse integrados y ser calculables, navegados e interconectados con fuentes de información relacionada que potencialicen su uso.

A lo largo del presente documento se mostrarán las diferentes capacidades de las tecnologías empleadas como recurso tecnológico para la aplicación del modelo semántico que ayudará a resolver los problemas de integración, cálculo, navegabilidad e interoperabilidad de datos financieros previamente descritos. Cada una de las tecnologías fue seleccionada en base a un extenso y detallado estudio acerca de sus propiedades mediante el análisis de la literatura y son reunidas enfocándose en el cumplimiento de los objetivos que se plantean en la sección siguiente.

Por otra parte, el conjunto de datos financieros generado con base en el modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009) que se presenta, permitirá validar los resultados obtenidos por la investigación realizada en este trabajo de tesis con el objetivo de comprobar las hipótesis que se plantean en el Capítulo 3 del presente documento.

1.3 Objetivo general y objetivos específicos

Para que se cumpla el objetivo general, es necesario que lleve a cabo el cumplimiento de cada uno de los objetivos específicos que como consecuencia permitirán dar respuesta a las hipótesis que se plantean en este trabajo de tesis. Por ende, tanto el objetivo general como los objetivos específicos son descritos en los apartados siguientes de esta sección.

1.3.1 Objetivo General

La definición de un modelo semántico que se inspira en los principios de Linked Data (T Berners-Lee, 2009) para la generación de una base de conocimientos financieros que sirva para proporcionar una alternativa de solución a las necesidades o limitaciones actuales concernientes a los Estados financieros basados en XBRL publicados por parte de las empresas, es, en términos generales el objetivo de este trabajo de tesis doctoral. Tales necesidades consisten en proporcionar a los datos extraídos de los Estados financieros XBRL, una semántica que permita su integración para ser navegables, facilitando el acceso a sus datos mediante el manejo de protocolos asociados a Internet como el HTTP para la navegación e interconexión con otras fuentes de datos, así como proveerles la capacidad para la realización de procesos de inferencia y el procesamiento de cálculos para su análisis fundamental. En este contexto, el análisis fundamental consiste en la identificación de los aspectos relevantes que ayuden a evaluar el valor de los Estados financieros de una empresa y sirvan de apoyo a la toma de decisiones (Ou & Penman, 1989).

1.3.2 Objetivos Específicos

La definición y la organización de los objetivos específicos es importante para el cumplimiento del objetivo general planteado, a continuación, se proporcionan los objetivos específicos a seguir durante esta investigación:

- Identificar y adquirir los Estados financieros que sean relevantes para el desarrollo de la investigación, a partir de fuentes de información disponibles en la Web.
- Analizar las principales normas para la publicación de Estados financieros, US-GAAP e IFRS.
- Definir las taxonomías financieras correspondientes a la norma de publicación de Estados financieros a implementar en la investigación.
- Analizar y diseñar un modelo conceptual basado en Linked Data para la transformación y representación semántica de datos financieros.

- Extraer y transformar de manera semántica los datos contenidos en los Estados financieros adquiridos. Esta transformación se apoya en el modelo semántico diseñado y consiste en la conversión de los datos financieros extraídos hacia tripletas RDF. Una transformación de este tipo permite combinar, exponer y compartir datos estructurados y semiestructurados de diferentes fuentes a través de múltiples aplicaciones (Cyganiak, Wood, & Lanthaler, 2014).
- Generar el conjunto de datos semántico que serialice y almacene los datos transformados manteniendo la definición del modelo conceptual.
- Validar de la calidad en la estructura de los datos financieros mediante la intervención de expertos en finanzas y contabilidad para validar los resultados obtenidos.
- Comprobar la utilidad de los datos para el cálculo de ratios financieros adicionales mediantes consultas basadas en SPARQL, que apoyen a la toma de decisiones tanto automatizada como por parte de las personas. Tal comprobación será verificada por un caso de estudio financiero y la intervención de expertos en finanzas y contabilidad para validar los resultados obtenidos.
- Comprobar la interconexión de datos con fuentes de información externas.
- Ofrecer un marco de trabajo común para la representación y cálculo de indicadores financieros.

Cada uno de los objetivos específicos mencionados tiene la finalidad de servir de instrumento para finalizar de manera exitosa la investigación de esta tesis, requieren del uso de tecnologías semánticas y del apoyo de especialistas en finanzas y contabilidad.

1.4 Organización de la tesis doctoral

Este trabajo de tesis inicia proporcionando una descripción general del entorno en el que se identifica la problemática a resolver, seguido de la motivación de la investigación, así como los objetivos generales y específicos. Posteriormente, se proporciona el estado del arte que inicia con información acerca del análisis fundamental financiero, los Estados financieros y las normas internacionales para su divulgación, el lenguaje XBRL y las tecnologías semánticas incluyendo Linked Data. Además, en el estado del arte se proporciona información de los trabajos e investigaciones relacionadas con el tema que se aborda en esta investigación. En el tercer capítulo se plantean las hipótesis a resolver y su proceso de validación, posteriormente en el Capítulo 4, se describen los modelos

semánticos de datos financieros, su conceptualización, sus características y su proceso de transición para la generación de una base de conocimientos financieros inspirada en los principios de Linked Data (T Berners-Lee, 2009). En capítulo siguiente, se describe la base de conocimientos generada a partir de los modelos semánticos presentados en el cuarto capítulo. El Capítulo 6, incluye información detallada de los experimentos que permiten la validación de las hipótesis planteadas, y en el Capítulo 7, se proporcionan las conclusiones generales del presente trabajo de tesis doctoral, así como sus líneas futuras de investigación.

Capítulo 2

Estado del arte

Resumen. Conocer información referente a los Estados financieros publicados por parte de las empresas, así como las tecnologías involucradas en ello, tiene una importancia significativa para esta investigación. La misma importancia tiene el recopilar información acerca de las tecnologías semánticas, modelos semánticos de datos e iniciativas de investigación y desarrollo realizadas por otros autores, y que están relacionadas con este trabajo de tesis. Toda esta información teórica y técnica es presentada en el actual capítulo, el cual concluye con una comparación entre los trabajos relacionados y la investigación correspondiente a este trabajo de tesis. Esta comparación incluye aspectos que permiten identificar semejanzas, diferencias y los posibles aportes que tiene este trabajo de tesis con los trabajos relacionados ya existentes.

2. Estado del arte

Con el propósito de establecer las bases teóricas y técnicas, así como identificar y conocer los resultados que en sus obras otros autores han obtenido, en el presente capítulo, se narran todos los conceptos, tecnologías e investigaciones relacionadas con la investigación presentada en este trabajo de tesis.

2.1 Análisis fundamental financiero

Durante el análisis de los Estados financieros se identifican aspectos relevantes para el apoyo en la toma de decisiones. Desarrollar un análisis ayuda a evaluar el valor de los Estados financieros de una empresa (J. A. Ou & Penman, 1989), complementando esta aseveración (Lev & Thiagarajan, 1993) mencionan que el análisis fundamental tiene por objeto determinar el valor de los títulos privados mediante un examen cuidadoso de factores clave/valor, tales como: los ingresos, las inversiones, el riesgo, el crecimiento y la posición competitiva, entre otros. Determinar el rendimiento de la empresa mediante un conjunto de medidas financieras también llamadas ratios financieros, es un problema interesante y desafiante para muchos investigadores y profesionales. La identificación de los ratios financieros que ayudan a predecir con exactitud el rendimiento de una empresa es de gran interés para cualquier tomador de decisiones (Delen, Kuzey, & Uyar, 2013). Por lo tanto, el análisis fundamental consiste en determinar la salud de las empresas mediante el análisis de sus ratios financieros históricos y actuales con el objetivo de realizar previsiones financieras, situación que se aplica al modelo semántico contemplado en este trabajo ya que proporciona una base de datos de conocimientos financieros computable para el análisis fundamental financiero.

2.1.1 Ratios financieros

Un ratio financiero es un indicador esencial para conocer la situación económica de la empresa (Bliss, 1923), y se expresa por medio de una fórmula matemática específica (generalmente simple). Los ratios financieros proporcionan información que beneficia a los interesados en la toma de decisiones empresariales incluyendo: propietarios, banqueros, inversores, consultores, gobiernos y muchos más. Además, ayudan a determinar la magnitud y la dirección de los cambios en la empresa durante un período de tiempo.

Básicamente, los ratios financieros se dividen en cuatro grupos principales (Montero & Fernández-Aviles, 2010):

1. **Ratios de liquidez:** evalúan la capacidad de la empresa para cumplir con sus compromisos de corto plazo.
2. **Indicadores de gestión o actividad:** son medidas que utilizan el activo y las ventas netas en comparación con el total de activos, los activos fijos tangibles, activos corrientes o elementos que pertenecen a ellas.
3. **Crédito, deuda o apalancamiento de calificaciones:** son ratios que relacionan los recursos y compromisos.
4. **Indicadores de rentabilidad:** miden la capacidad de la empresa para generar riqueza (económica y financiera).

Uno de los propósitos de esta tesis consiste en ofrecer capacidades de análisis fundamental financiero que sirvan de apoyo para la toma de decisiones tanto automatizada como por parte de las personas. Por tal motivo, en los siguientes tres grupos se examina un subconjunto de indicadores financieros derivado de los cuatro grupos de ratios descritos en la lista anterior y que son útiles para el análisis contable de una empresa (Kimmel, Weygandt, & Kieso, 2010; Montero & Fernández-Aviles, 2010):

1. **Ratios de rotación:** también son llamados ratios de eficiencia o actividad, proporcionan información sobre la capacidad de la administración para controlar los gastos y obtener un rendimiento sobre los recursos asignados a la empresa. Estos ratios miden los rendimientos originados por los activos que una entidad obtiene en un período determinado, se utilizan como complemento de los ratios de rentabilidad y su resultado se mide en número de veces. No obstante, para que realmente sean representativos, es necesario utilizar valores promedio, ya que es normal que sus componentes fluctúen a lo largo del ejercicio e interesan aquellos valores que son elevados, pues indican que la empresa genera mayores ventas con menos inversión. Los Ratios de rotación se clasifican de la siguiente manera, Rotación del activo total, Rotación del activo no corriente, Rotación del activo corriente, Rotación de los clientes, Rotación de almacén o Rotación de *Stocks* y Rotación del capital circulante.
2. **Ratios de liquidez:** estos ratios corresponden a la capacidad de una empresa para convertir sus activos en caja, midiendo la solvencia que esta tiene en corto plazo para cumplir con sus obligaciones actuales. Los Ratios de liquidez se clasifican en, Capital de trabajo, Capital de trabajo neto sobre el total de los activos, Capital de

trabajo neto sobre deudas a corto plazo, Test ácido, Días de medición del intervalo de tiempo, y Razón de efectivo.

3. **Ratios de solvencia:** son una medida de la viabilidad financiera a largo plazo de una empresa, su finalidad es la de diagnosticar si una entidad tiene problemas para atender sus compromisos, entre ellos, la liquidación de sus deudas en la cuantía y plazos pactados. Para evaluar la política financiera de la empresa, es decir, la cantidad y la calidad de la deuda, se utilizan los siguientes ratios, Ratio de endeudamiento y Ratio de calidad de deuda.

Los ratios mencionados previamente para el análisis contable de una empresa, son calculables con el modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009) que se presenta en esta tesis. Las Tablas 1, 2 y 3, presentan la información correspondiente a cada ratio, clasificándolos por grupo e indicando el nombre del ratio, su fórmula como indicador financiero, una breve definición y su utilidad para las empresas. Cabe mencionar que el nombre de los ratios y sus fórmulas, son proporcionados tanto en idioma Español como en Inglés con la finalidad de identificarlos con su correspondiente equivalencia con las taxonomías financieras basadas en la norma US-GAAP que se describen en el Capítulo 4 de este trabajo de tesis. Adicionalmente, la Tabla 3 contiene las abreviaturas que simplifican la comprensión de las fórmulas descritas en cada Tabla.

INDICADORES FINANCIEROS - RATIOS DE ROTACIÓN (TURNOVER RATIOS)			
RATIO	**FÓRMULA**	**DEFINICIÓN**	**ÚTILIDAD PARA LAS EMPRESAS**
Rotación del Activo Total, *(Total Assets Turnover).*	Ventas Netas / Activos Totales *Net Sales / Total Assets*	Mide el número de veces que los ingresos por ventas cubren las inversiones (Activo Total) de la empresa, o lo que es lo mismo, el rendimiento que proporcionan los Activos totales (Ventas que se producen con la Inversión realizada).	Sirve para conocer cuántas ventas se producen gracias a los Activos de los que la empresa dispone. En función de esto, se aumentarán, mantendrán o disminuirán, los activos de la empresa.
Rotación del Activo no Corriente, *(Non-Current Assets Turnover).*	Ventas Netas / Activos No Corrientes *Net Sales / Non-Current Assets*	Se obtiene dividiendo las ventas entre el activo no corriente formado por el inmovilizado material e intangible, inversiones y créditos financieros y las inversiones inmobiliarias.	Refleja las veces que se ha utilizado el activo no corriente en la obtención de ingresos, es un índice de la eficiencia en la gestión de los bienes del activo no corriente y es deseable que su valor sea lo más elevado posible.
Rotación del activo corriente, *(Current Assets Turnover).*	Ventas Netas / Activos Corrientes *Net Sales / Current Assets*	Se determina como el cociente entre las Ventas netas y el Activo corriente. El Activo corriente está formado por las existencias, deudores y derechos de cobro a corto plazo, inversiones y créditos financieros con vencimiento en menos de un año y la tesorería.	Es útil para conocer cuántas ventas se producen gracias a los Activos corrientes o circulantes de los que la empresa dispone. En función de esto, se aumentarán, mantendrán o disminuirán, los Activos corrientes o circulantes de dicha empresa.
Rotación de los Clientes *(Turnover of Customers).*	Ventas Netas / DDC *Net Sales / DDC*	Mide las rotaciones de las cuentas por cobrar y se utiliza para evaluar las condiciones de pago que la empresa concede a sus clientes.	Mide el número promedio de veces al año que se cobran las cuentas a clientes, es decir, la frecuencia de recuperación de las cuentas por cobrar. Un valor elevado de este ratio indica que la empresa no tarda mucho tiempo en recuperar sus ventas.
Rotación de Almacén o Stocks, *(Stock Rotation).*	Ventas Netas / Inventario *Net Sales / Inventory*	Indica el número de veces que los ingresos por ventas cubren la inversión en existencias e indica el número de veces que se renueva el inventario o stock en un período económico.	Es una de las métricas de eficiencia de la cadena suministro más utilizadas. Interesa que su valor sea lo más alto posible; si fuera bajo, indicaría que la empresa retiene en su almacén existencias consecuencia de menores ventas, aspecto que habría que analizar al detalle.
Rotación del Capital Circulante, *(Capital Turnover).*	Ventas Netas / CC = Ventas Netas / Fondo de Rotación Existente Operacional *Net Sales / WC = Net Sales / Operational Existing Revolving Fund*	Indica el número de veces que las ventas cubren el capital circulante o corriente, o, lo que es lo mismo, la inversión en el ciclo de explotación o fondo de rotación existente operacional.	Interesa que su valor sea alto y se complementa con una comparación con el valor representativo del sector. También se llama rotación de la equidad y permite conocer el ciclo de la obtención de capital, la realización de inversiones y la devolución del capital para su fuente de beneficios generados.

Tabla 1. Ratios de rotación (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

INDICADORES FINANCIEROS - RATIOS DE LIQUIDEZ (LIQUIDITY RATIOS)			
RATIO	**FÓRMULA**	**DEFINICIÓN**	**ÚTILIDAD PARA LAS EMPRESAS**
Capital de trabajo, *(Working Capital).*	CA - CL	Busca garantizar las operaciones de la empresa, es la diferencia entre el Activo corriente o circulante (duración de no más de un ejercicio económico) y el Pasivo corriente o exigible a corto plazo.	Si el resultado es positivo, viabiliza la generación de inversión, y si es negativo, indica la necesidad de buscar financiamiento. El financiamiento debe servir para que el Capital de trabajo pase de negativo a positivo, además de pagar las deudas, los intereses y permita generar utilidad.
Capital de trabajo Neto sobre el Total de Activos. *(Net Working Capital Over Total Assets)*	(CA – CL) / TA	Mide la relación del Capital de trabajo que corresponde al dinero que posee la empresa para trabajar. Que dicho de otro modo, es el capital que tiene una empresa tras haber pagado sus deudas en el corto plazo con sus activos disponibles.	El nivel óptimo es que el valor sea mayor que 0, una razón baja indica niveles de liquidez bajos (no tiene un número adecuado de Activos circulantes). La interpretación del resultado de este ratio dependerá del sector en el que opera la empresa.
Capital de trabajo Neto Sobre Deudas a Corto Plazo, *(Net working Capital on Short-Term Debt).*	(CA - CL) / CL	Consiste en el cociente entre la diferencia entre el Activo y el Pasivo corriente o circulante (menos de un ejercicio económico) y el pasivo corriente o exigible a corto plazo.	Su nivel es óptimo cuando el valor es cercano a 0,5. Si es menor que 0,5 es posible que la empresa tenga problemas para cumplir con sus deudas a corto plazo, aunque convierta en dinero todos sus activos.
Test Ácido, *(Acid Test).*	(CA – Inventario) / CL (CA – Inventory) / CL	Revela la capacidad de la empresa para cancelar sus obligaciones corrientes pero sin contar con la venta de sus existencias (inventarios), consiste en el cociente dado por la diferencia entre el Activo corriente o circulante y la Cuenta de Inventario y el Pasivo Corriente o exigible a corto plazo.	Revela la capacidad de la empresa para cancelar sus obligaciones corrientes pero sin contar con la venta de sus existencias. Si es menor a 1, la empresa podría suspender sus pagos con terceros y si es mayor que 1 indica la posibilidad de que haya un exceso de liquidez. Los valores óptimos se encuentran entre 0,5 y 1.
Días de medición del intervalo de tiempo, *(Day Time Interval Measurement).*	(CA - CM) * 365	Corresponde al cociente entre los activos circulantes y el costo de la mercadería, multiplicado por 365 días.	Mide el número de días en cual la empresa puede seguir operando, si, por cualquier motivo, estuviese paralizada en sus actividades cotidianas.
Razón corriente, *(Current Ratio).*	CA / CL	La Razón corriente permite determinar el índice de liquidez de una empresa, esto significa que el resultado obtenido es determina la capacidad de pago que tiene una empresa.	Es la principal medida de liquidez, Entre mayor sea la razón resultante, mayor solvencia y capacidad de pago tendrá la empresa para saldar sus deudas.
Razón de Efectivo, *(Cash Ratio).*	EF / CL	Es la razón que relaciona las inversiones financieras temporales que una empresa puede convertir en efectivo en 1 ó 2 días. Esto excluye aquellas cuentas bancarias que no sean de libre disposición por estar afectas a garantía.	El nivel óptimo es obtener valores de 0,3. Esto significa que por cada unidad monetaria que se adeuda, se poseen 0,3 unidades monetarias de efectivo en 2 ó 3 días.

Tabla 2. Ratios de liquidez (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

INDICADORES FINANCIEROS - RATIOS DE SOLVENCIA (SOLVENCY RATIOS)			
RATIO	**FÓRMULA**	**DEFINICIÓN**	**ÚTILIDAD PARA LAS EMPRESAS**
Coeficiente de endeudamiento, *(Debt to Equity Ratio)*	RA / RP	Este ratio mide la política de financiación que emplea la entidad, comúnmente se denomina apalancamiento financiero. Se calcula como el cociente entre los recursos ajenos y los propios, reflejando la relación entre ambos. Cabe mencionar que no existe un valor adecuado o idóneo, puesto que depende de cada entidad.	Su interpretación es bastante sencilla: a) Si es igual a la unidad, indica que la empresa emplea la misma cantidad de recursos ajenos que propios; b) Si es menor que la unidad, la entidad tiene mayor proporción de financiación propia; y c) Si es mayor a la unidad, la entidad se financia más con recursos obtenidos de terceros que con propios. Por último, hay que resaltar que la tenencia de deuda no siempre es perjudicial. De hecho, existe deuda no remunerada (financiación espontánea) y es posible que tenga menor coste de recursos propios.
Ratio de calidad de la deuda, *(Ratio of Debt Quality).*	RACP / RA	Es el cociente obtenido entre los recursos ajenos a la empresa que solicita en el corto plazo (menos de un ejercicio económico) y los recursos ajenos totales de la misma.	Se dice que las deudas de la empresa son de mejor o peor calidad en función de su plazo, así cuanta más financiación a corto plazo se tenga, se interpreta que es de peor calidad, pues, en un breve plazo (no más de un ejercicio) la empresa tendrá que desprenderse de recursos económicos para atender a su pago. Cuanto más pequeño sea este ratio significa que la deuda es de mejor calidad.
Razón de deuda, *(Debt ratio)*	((Pasivos totales / TA)*100) *((Total Liabilities / TA)*100)*	Permite establecer el grado de participación de los acreedores, en los activos de la empresa. Mide la proporción de la inversión de la empresa que ha sido financiada por deuda, por lo cual, se acostumbra presentar su resultado en forma de porcentaje.	Se considera que un endeudamiento del 60% es manejable, es decir, que de cada 100.00 USD que la empresa tiene en sus activos se adeudan 60.00 USD. Esta razón de endeudamiento es 0,6 e indica que el 60% del total de la inversión (Activos Totales) ha sido financiada con recursos de terceros (endeudamiento).

ABREVIATURAS (ABBREVIATIONS)

CA: Activo Corriente o Activo Circulante	**EF:** Efectivo	**CA:** Current Assets	**EF:** Cash
CL: Pasivo Circulante o Pasivo exigible a corto plazo	**RA:** Préstamos o Recursos Ajenos	**CL:** Current Liabilities	**RA:** Borrowings
CM: Costo de los Materiales	**RACP:** Recursos ajenos a corto plazo	**CM:** Cost of Materials	**RACP:** Short-Term Borrowings
CC: Capital Circulante	**RP:** Equidad o Recursos Propios	**WC:** Working Capital	**RP:** Equity
DDC: Derechos de cobro comerciales	**TA:** Activo Total	**DDC:** Trade Receivables	**TA:** Total Assets

Tabla 3. Ratios de Solvencia (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

2.1.2 Estados financieros

Un estado financiero es un documento que recoge contablemente la situación financiera en la que se encuentra una organización en un determinado momento, este momento suele ser a final de un ejercicio anual, de un semestre o de un trimestre. Siguiendo este contexto, instituciones gubernamentales como la U.S. SEC que, a través del Sistema EDGAR (*Electronic Data Gathering Analysis and Retrieval System,* Sistema de Recopilación, Análisis y Recuperación de Datos Electrónicos), se encargan de almacenar y publicar en la Web los Estados financieros de las empresas en Estados Unidos, mediante conjuntos de informes o *filings* basados en el estándar XBRL. Estos *filings* cumplen con ciertas reglas y formatos, entre los que vale la pena mencionar la asignación de claves para definir sus periodos de publicación siendo: 10-K el formato para la presentación de informes anuales, N-SAR para presentar informes semestrales y 10-Q para los informes trimestrales (U.S. SEC., 2013). Para la presente investigación, el formato 10-Q resulta muy conveniente, ya que proporciona una visión continua de la posición financiera de la empresa durante el transcurso del año.

Hay dos categorías principales de datos para el análisis y el desarrollo de cálculos financieros, estos son: datos en tiempo real y datos de divulgación pública. Estos últimos se dividen en los siguientes tres tipos de Estados financieros (Penman, 2009):

1. **Hoja de Balance *(Balance Sheet)*:** es un documento contable que refleja la situación financiera de una empresa en un tiempo determinado, constatando, sistematizando y valorando sus activos, pasivos y su patrimonio neto. En el caso de las sociedades que cotizan en Bolsa, deben presentarlo al final de cada trimestre.
2. **Cuenta de resultados *(Income Statement)*:** es un documento contable que muestra los resultados de las operaciones (utilidad, pérdidas y excedentes) de una entidad durante un periodo determinado. Se presenta la situación financiera de una empresa a una fecha determinada, tomando como parámetros los ingresos y los gastos.
3. **Estado de flujos de efectivo *(Cash Flow)*:** el estado de flujos de efectivo proporciona los datos relativos a todas las entradas de efectivo que una empresa recibe de sus operaciones en curso y fuentes de inversión externa, así como todos los flujos de efectivo relacionados con actividades comerciales y las inversiones durante un trimestre dado.

La transformación de los datos de los Estados financieros en forma semántica, abre una alternativa para la generación de una base de conocimientos basada en Linked Data que permita su integración e interoperabilidad así como el cálculo de los ratios financieros. Valiéndose del hecho de que por sí mismos, los conceptos de los Estados financieros publicados en la Web son aprovechables para consultar y recuperar cifras útiles, se hace factible utilizar la semántica para definir ratios financieros a través de la generación de las taxonomías financieras (véase Capítulo 4) que forman parte indispensable del modelo semántico inspirado en Linked Data de la presente tesis.

2.2 Normas internacionales de información financiera

El Comité de Normas Internacionales de Contabilidad *(IASC, International Accounting Standards Committee),* establecido en 1973 y actualmente llamado Consejo de Normas Internacionales de Contabilidad *(IASB, International Accounting Standards Board)*[4], tiene como objetivo lograr la uniformidad en las normas contables utilizadas por las empresas y otras organizaciones para la información financiera en todo el mundo (IASB-IFRS, 2014). Se considera que los beneficios de la adopción de normas internacionales de contabilidad son los siguientes. En primer lugar, se mejora la capacidad de los inversores para tomar decisiones financieras informadas, eliminando la confusión derivada de las diferentes medidas de la posición financiera y el rendimiento de los distintos países, lo que conduce a una reducción del riesgo para los inversores y un menor costo de capital para las empresas. En segundo lugar, se reducen los costes derivados de la presentación de informes múltiples. En tercer lugar, se fomenta la inversión internacional. Por último, se conduce hacia una asignación más eficiente de los ahorros de todo el mundo (Street, Gray, & Bryant, 1999).

Las Normas Internacionales de Contabilidad originales eran en su mayoría de carácter descriptivo y contenían muchos tratamientos alternativos. Debido a esta flexibilidad y una continua falta de comparabilidad entre los países, las normas fueron objeto de fuertes críticas a finales de 1980. En respuesta a estas críticas, el IASC inició el Proyecto de Comparabilidad en 1987. Las normas revisadas, las cuales entraron en vigor en 1995, redujeron sustancialmente los tratamientos alternativos y el aumento de los requisitos de divulgación (Nobes & Parker, 2008). En Julio de 1995, la IASC y la Organización Internacional de Comisiones de Valores (*IOSCO, Organization of Securities Commission,* OICV,

[4]IFRS-IASB: http://www.ifrs.org/Pages/default.aspx

Organización Internacional de Comisiones de Valores)[5], acordaron una lista de cuestiones contables que era necesario abordar para obtener el refrendo de los estándares del IOSCO. El subsiguiente *Core Standards Project,* condujo nuevamente a las revisiones sustanciales de las Normas Internacionales de Contabilidad (NIC) o IAS *(International Accounting Standards),* por sus siglas en Inglés.

En Mayo del año 2000, el IASC recibió de la IOSCO la aprobación sujeta a la *"reconciliación cuando sea necesaria para hacer frente a las cuestiones pendientes y correspondientes a los niveles nacional o regional"* (IOSCO, 2000). El *Core Standards Project,* ganó un mayor reconocimiento de las NIC en todo el mundo. Por ejemplo, el Parlamento Europeo emitió el reglamento (1606/2002/CE) en el que solicitó que para el año 2005, todas las sociedades cotizadas de la Unión Europea prepararen sus Estados financieros consolidados con base en las Normas Internacionales de Contabilidad. Aunque antes del año 2005, en una serie de países, entre ellos Austria, Bélgica, Francia, Alemania, Italia y Suiza, las compañías ya estaban autorizadas a formular cuentas anuales consolidadas bajo las NIIF (Normas Internacionales de Información Financiera) o IFRS *(International Financial Reporting Standards)* por sus siglas en Inglés, así como los PCGA de EE.UU (Principios de Contabilidad Generalmente Aceptados), o US-GAAP *(US-Generally Accepted Accounting Principles)* también por sus siglas en Inglés, desarrollados por el FASB (*Financial Accounting Standards Board,* Consejo de Normas de Contabilidad Financiera)[6] (Van-Tendeloo & Vanstraelen, 2005).

2.2.1 Normas internacionales US-GAAP e IFRS

Con la finalización exitosa de las normas del IASC, las IFRS y los US-GAAP fueron colocados como los dos marcos de información financiera preeminentes a nivel mundial. Sin embargo, los US-GAAP fueron aceptados ampliamente como el conjunto internacional de normas para garantizar Estados financieros de alta calidad. No sólo en los EE.UU., sino también en otros lugares, lo que ha dado lugar a un debate sobre la (relativa) la calidad de ambos regímenes (Der-Meulen, Gaeremynck, & Willekens, 2007). Mientras que los defensores del IFRS argumentan que su calidad ha mejorado considerablemente con el tiempo, especialmente con la realización del Proyecto de Comparabilidad/mejora, ahora están relativamente cerca de los US-GAAP con sólo pequeñas diferencias restantes y dan revelaciones suficientes que permiten a los inversores sacar sus propias conclusiones en

[5]IOSCO: http://www.iosco.org/
[6]US-GAAP: http://www.fasb.org/home

caso de divergencia. Por otra parte, los opositores argumentan que las diferencias de IFRS son aún considerables entre los dos estándares y mantienen que los US-GAAP son aún de mejor calidad ya que los IFRS presentan información menos detallada (Leuz, 2003).

Acedo, (2006) proporciona un resumen acerca de las principales diferencias entre IFRS y US-GAAP, lo que podría suponer que no es viable la convergencia de ambas normas. Tal suposición es complementada por la conclusión de (Gordon, Jorgensen, & Linthicum, 2008) en la que indican que los Estados financieros publicados bajo US-GAAP exhiben un mayor contenido informativo en relación con IFRS, lo que sugiere que una mediación de IFRS a US-GAAP da lugar a Estados financieros menos útiles para la valoración de las acciones de las empresa. Sin embargo, el propio (Acedo-Peñalva, 2006) y (Laguna & Romero, 2009), argumentan que en Septiembre del año 2002, el FASB y el IASB alcanzaron el acuerdo *Norwarlk* (FASB, 2002), el cual supuso un hito importante que indica que ambos organismos reguladores se comprometieron a desarrollar normas de gran calidad que pudieran ser utilizadas tanto en el *reporting* nacional como en el internacional. Adicionalmente, (Laguna & Romero, 2009) mencionan que en el acuerdo de intenciones firmado en el año 2006, tanto el FASB como el IASB reconocieron la importancia de cumplir con el calendario trazado para eliminar la obligación de reconciliar las cifras hacia los US-GAAP, como muy tarde en el año 2009, para aquellas sociedades extranjeras que coticen en Estados Unidos y utilicen IFRS en sus Estados financieros. Para ello, se estableció un programa de convergencia para el periodo 2006-2008, distinguiendo ciertas areas incluidas en el programa de convergencia a corto plazo de otros proyectos conjuntos.

Considerándo los trababajos descritos en esta sección, es importante resaltar que el tema sobre las normas IFRS, US-GAAP y su convergencia, es un tema muy amplio y con investigaciones en progreso, que no corresponden directamente a esta investigación. Sin embargo, resulta conveniente para este trabajo de tesis, hacer uso de la norma US-GAAP, dadas las afirmaciones de algnos autores como (Gordon et al., 2008), que mencionan que los informes financierons basados en la norma US-GAAP ofrecen información más detallada y que dicha norma cuenta con una mayor aceptacion por parte de las principales empresas a nivel internacional.

2.3 El lenguaje XBRL

El inicio del *eXtensible Business Markup Language* o XBRL se remonta al mes de Abril de 1998, cuando el contador público y auditor Charles Hoffman comenzó a desarrollar prototipos para la información financiera utilizando el lenguaje XML. El Instituto

Americano de Contadores Públicos Certificados *(AICPA, American Institute of Certified Public Accountants)*[7] apoyó y financió la iniciativa de Hoffman y en junio de 1999, él y varios de sus colaboradores crearon un plan de negocio para los Estados financieros basados en XML, llamado XFRML (*eXtensible Financial Reporting Markup Language,* Lenguaje Extensible de Marcado para la Información Financiera) (Wu & Vasarhelyi, 2004) al que posteriormente se le cambió nombre y en Abril del año 2000, se presentó públicamente como la primer especificación XBRL (Debreceny & Gray, 2001).

XBRL es una versión flexible de XML, desarrollada específicamente para satisfacer las exigencias de la información financiera y empresarial, que permite el intercambio de esta (Balance general, Cuenta de pérdidas y ganancias, Estado de flujo de efectivo, entre otros) utilizando los mecanismos más habituales de Internet, como la Web y el correo electrónico. Aplica etiquetas identificativas únicas a los distintos elementos que componen la información financiera. Estas etiquetas son algo más que solamente identificativas, pues proporcionan una amplia gama de información sobre el elemento, por ejemplo, el porcentaje o fracción. XBRL, permite que se utilicen etiquetas en cualquier idioma, así como referencias contables u otra información complementaria (Hoffman & Van-Egmond, 2012).

2.3.1 XML como lenguaje universal y abierto

El *eXtensible Markup Language* o XML fue propuesto por el *World Wide Web Consortium* (Consorcio de la W3C) para proporcionar información estructurada a la Web y es utilizable en muchos dominios diferentes proporcionando una forma estándar para buscar, visualizar, manipular e intercambiar datos en la Web (Bray et al., 1998; Zisman, 2000).

XML proporciona un conjunto de elementos que se utilizan para definir los tipos de documentos, estos elementos se conocen como DTD (*Document Type Definitions,* Definición de Tipo de Documento). Un DTD contiene un conjunto de reglas para controlar cómo están estructurados los documentos XML, que elementos presentan y las relaciones entre estos elementos (Papakonstantinou & Vianu, 2000). Siguiendo este contexto, XML es un lenguaje ideal para la publicación de documentos semiestructurados tales como manuales, catálogos, informes, patentes, Estados financieros y documentos de investigación. Además, XML es muy útil para facilitar la generación y gestión de metadatos (los llamados metadatos de las etiquetas) que proporcionan un significado adicional a la información

[7] AICPA: http://www.aicpa.org/Pages/default.aspx

ordinaria contenida en este tipo de documentos, de manera que las aplicaciones informáticas que consumen esta información sean capaces de entender su significado (Zisman, 2000).

En un documento XML, la información ordinaria se acota entre etiquetas para dotarlas de significado adicional. Un ejemplo de esto se muestra en la Tabla 4, en la que se observa como una aplicación informática es capaz de interpretar los metadatos de las etiquetas contenidas en un documento XML para entender el nombre de la persona *"Michael M. Miller"* y distinguirlo separadamente de su dirección postal y de esta forma automatizar el procesamiento de estos datos.

INFORMACIÓN ORDINARIA	METADATOS INFORMACIÓN ACERCA DE LA INFORMACIÓN
Michael M. Miller **300 Boylston Ave E** **Seattle, WA 98102** **USA** **206-684-3020** **Birthdate: 19-October 58**	<name>**Michael M. Miller**</name> <address> <street>**300 Boylston Ave E**</street> <city>**Seattle**</city> <state>**Washington**</state> <country>**United States of America**</country> <zipcode>**98102**</zipcode> </address> <telephone>**206-684-3020**</telephone> <birthdate>**1958-10-19**</birthdate>

Tabla 4. Ejemplo de información ordinaria y metadatos

Desde su creación en 1998, XML ha servido como base para la construcción de otros lenguajes según diversos aspectos (XBRL-España, 2006):

- **Orientados al intercambio y extracción de información:** SOAP, WSDL, XQuery, XPath, SAX, DOM.
- **Orientados a formar Vocabularios específicos de negocio:** MathML, MusicML, OTA, HL7, XBRL.
- **Orientados al formato o presentación de la información:** XHTML, XForms, WML, SVG.
- **Orientados para tratar y transformar el propio XML:** XSLT, XSL-FO, XML-Schema, RelaxNG, XLink, XPointer

Dadas sus características, XML es un lenguaje universal y abierto que permite el intercambio de información estructurada entre diferentes plataformas y mejora la potencia y la flexibilidad de las aplicaciones Web y otros paquetes de software de negocios (Harold, 2004; W3C, 2011b). En el contexto de los lenguajes como XML, un formato abierto se

refiere a una especificación para almacenar datos digitales, la cual es publicada y patrocinada habitualmente por una organización de estándares abiertos y libre de restricciones legales y económicas de uso (ISO, 2008b).

2.3.2 XBRL como lenguaje estándar para la presentación de información financiera

XBRL se basa en XML porque su sintaxis mantiene un formato universal y abierto que permite que las definiciones de los metadatos a intercambiar sean definiciones estándar, es decir, que un término como *"Caja y depósito en Bancos Centrales"* por citar un ejemplo, signifique siempre lo mismo independientemente de las aplicaciones que utilicen dicho término (XBRL-España, 2006). Este es el pilar en el que se sustentan las taxonomías y diccionarios comunes de datos expresados en lenguaje XBRL (Hoffman & Van-Egmond, 2012).

Los elementos más importantes de XML son: a) La validación de la exactitud; b) La complejidad estructural; y c) Un conjunto de etiquetas extensible. Asociado con la definición principal de XML, hay una variedad de estándares asociados con el W3C incluyendo espacios de nombres *(Namespaces)* para la identificación única de dominios, apuntadores a recursos (XLink), hojas de estilo (XSL), formularios basados en Web (XForms), esquemas XML (XML-Schema), modelos abstractos de datos (XML-Infoset), identificación única de fragmentos de documentos XML (XML-Fragment) para la realización de vínculos (X-Point), una versión XML de HTML 4.0 (XHTML), y un XML para dispositivos móviles (Debreceny & Gray, 2001).

2.3.2.1 Expresión de reglas con XBRL

Para conseguir la extensibilidad y garantizar la unicidad en la definición de los conceptos, el etiquetado de información XML es insuficiente, por lo que es necesaria la adición de reglas a esa información. Las reglas dotan de semántica a la información expresada en las taxonomías, esto significa que en el lenguaje XBRL, adicionalmente a la definición de los conceptos a reportar, se expresan unos metadatos de los propios conceptos de la taxonomía, esto son, las relaciones existentes entre los conceptos, y se expresan mediante reglas con sintaxis XML, las llamadas *linkbases* (XBRL-España, 2006).

Los *Linkbases* (DeRose et al., 2000) son colecciones de enlaces XLink extendidos, que proporcionan más información sobre el significado de los conceptos mediante la expresión de las relaciones entre los propios conceptos (relaciones inter-concepto) y por la asociación

de conceptos para su documentación. Las taxonomías XBRL hacen uso de cinco tipos de *linkbases* las cuales son, *linkbases* de definición, de cálculo, de presentación, de etiquetas y de referencia. Cada uno de ellos se describe brevemente en la sección siguiente.

2.3.3 Estructura básica de los Estados financieros basados en XBRL

En XBRL existen dos secciones diferenciadas, la primera consiste en la especificación de taxonomías que definen los conceptos que se mostrarán en los Estados financieros junto con información adicional relacionada con los conceptos definidos. Por ejemplo, etiquetas (textos) en múltiples idiomas, relaciones de cálculo entre los conceptos, fórmulas que suponen reglas de validación y referencias a normas legales relacionadas con el concepto, por mencionar algunos. Las taxonomías XBRL son como el formulario y las reglas que permiten verificar el contenido del formulario. La segunda sección corresponde a los Informes XBRL en los que se identifican las taxonomías a emplear y se introduce la información usando los conceptos definidos previamente (XBRL-España, 2006).

Una taxonomía XBRL está constituida por la definición de un esquema XML (*XML Schema*) a través de un XSD *(XML Schema Definition)* (Biron et al., 2012; Thompson, 2012) y por los *linkbases* XLink contenidos o referenciados por ese esquema. Una taxonomía puede ser parte de un conjunto de taxonomías relacionadas llamado Conjunto de Taxonomías Detectables *(DTS, Discoverable Taxonomy Set)* (Engel et al., 2008). El XML Schema en una taxonomía, define los conceptos de la información a los que se les da un nombre y un tipo, como definiciones de elementos XML Schema. La Figura 1, muestra la estructura de las taxonomías que integran a los Estados financieros publicados en XBRL.

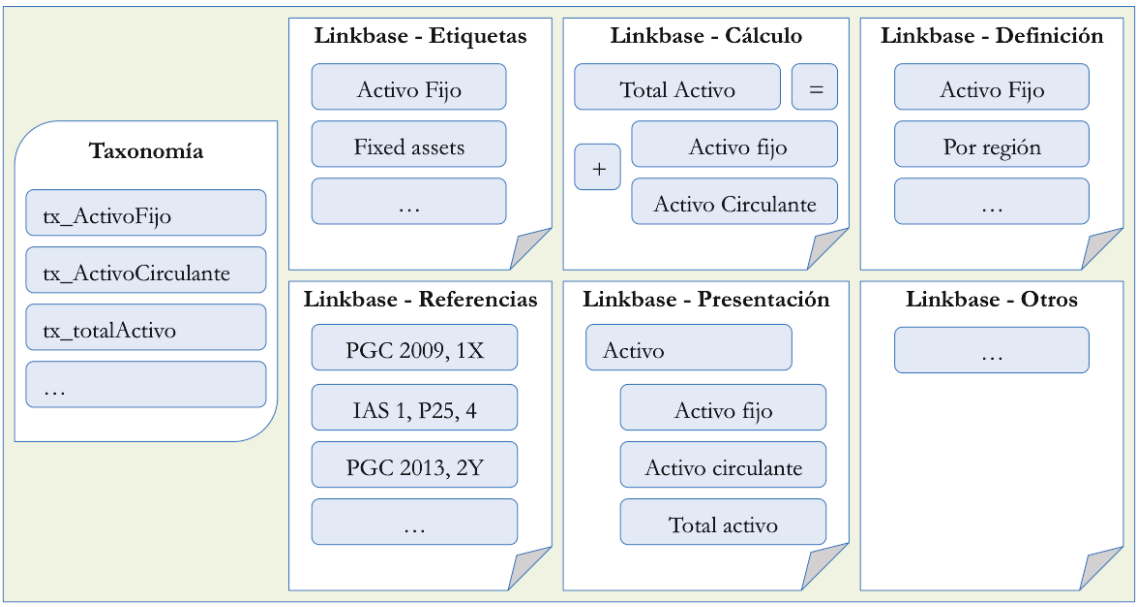


Figura 1. Estructura de las taxonomías que integran a los Estados financieros publicados en XBRL

Según la teoría de la comunicación, para que se intercambie un mensaje entre un emisor y un receptor, es necesaria la existencia de un código que sea conocido por los participantes. Este es el papel de las taxonomías XBRL, las cuales como ya se ha mencionado, están constituidas por esquemas XML (XBRL-España, 2006). Estos esquemas definen el conjunto de elementos que aparecen en los informes y la estructura de los mismos, tal conjunto se denomina diccionario de términos definidos que como se muestra en la Figura 1, tienen una función determinada la cual se describe brevemente a continuación (Engel et al., 2008; XBRL-España, 2006):

- **Linkbase de etiquetas:** son etiquetas o textos asociados a los elementos del diccionario que permiten ser utilizados en distintos idiomas y con distintos propósitos a la hora de construir representaciones de los informes. Los humanos interpretamos fácilmente que el dato corresponde al concepto que aparece en la misma fila, por ejemplo: *"Activo fijo 150,000 €"*.
- **Linkbase de referencias:** estas son referencias a textos legales o normativas que fundamentan la base legal del concepto a modelar. Estas referencias juegan un papel muy importante a la hora de aclarar la utilización de los conceptos cuando se van a crear los informes. Esta *linkbase* es de mucha utilidad a la hora de localizar los conceptos que deben usarse para elaborar informes XBRL que utilicen alguna taxonomía bajo un estándar para la publicación de información financiera como IFRS o US-GAAP.
- **Linkbase de presentación:** define las reglas para construir una representación del informe que se pretende modelar. Esta *linkbase* tiene un doble propósito, por un lado, sirve para que las herramientas de creación o visualización de taxonomías muestren el contenido de la misma de forma más amigable que una simple lista de conceptos y sirve de base para que las aplicaciones que formatean los informes de forma automática tengan un punto de partida por el que empezar a construir las plantillas que mostrarán los datos.
- **Linkbase de cálculo:** se encarga de establecer las reglas de cálculo, (operaciones aritméticas) entre los elementos de la taxonomía que permiten validar los informes XBRL.
- **Linkbase de definición:** son reglas adicionales que permiten establecer relaciones entre los elementos de una taxonomía que permitan explicar o documentar

relaciones, así como añadir ciertas reglas que son importantes para validar documentos XBRL.

Las *linkbases* son también extensibles. Nada en la especificación impide que se desarrollen *linkbases* propietarias para relacionar modelos de datos internos con elementos de taxonomías, esas *linkbases* deberán ser privadas, dado que no existe una especificación aprobada en el consorcio para que todos los procesadores XBRL las entiendan (Engel et al., 2008).

2.3.4 Transparencia financiera con XBRL

Diversas investigaciones acerca de la transparencia informativa indican que mientras mayor transparencia exista en los formatos de los Estados financieros, más fácil será para los usuarios detectar una Manipulación de Beneficios, sobre todo, en aquellos formatos que son más fáciles de procesar (Hirst & Hopkins, 1998). En relación con el argumento previo, los administradores a menudo cabildean para elegir formatos de divulgación menos transparentes (por ejemplo, la inclusión de los gastos y pasivos en las notas y los cambios de valor de mercado en la declaración de patrimonio de los accionistas, entre otros). Esto sugiere que los administradores pueden creer que hay un beneficio derivado de la limitación en la capacidad que tienen algunos usuarios para detectar una Manipulación de Beneficios (Hunton, Libby, & Mazza, 2006). Tal creencia es consistente con la perspectiva proporcionada por Fields, Lys, & Vincent (2001) en la que sugieren la probabilidad de que los administradores racionales no participan en la Manipulación de Beneficios por la ausencia de ganancias o beneficios esperados. Para que tal ausencia exista, se requiere que los usuarios involucrados con la información contable sean incapaces o no estén dispuestos a desentrañar sus efectos. Esto permite inferir que si las personas tienen la capacidad de detectar fácilmente los beneficios a obtener tras la interpretación de los Estados financieros, se reducirá el valor de la Manipulación de Beneficios lo que implica una mayor transparencia en la información ayudando a reducir la prevalencia en los intentos de la Manipulación de Beneficios a fin de mejorar la transparencia de la información.

Apegándose más al contexto empresarial, los usuarios que hacen uso de los datos financieros en general, han concebido que la dependencia de las empresas hacia los sistemas de reportes financieros actuales y un enfoque basado en los ingresos, han dado lugar a un mayor riesgo en el mercado por la manera en la que los administradores tratan de manejar las ganancias (Allen & Cote, 2005). De manera específica, estos usuarios

descubrieron que reportar ganancias trimestralmente se presta a tener una interpretación sospechosa disminuyendo la transparencia de la información, ya que estos y otros datos financieros no siempre son vistos como un reflejo de la rentabilidad a largo plazo de una empresa (Ahmadpour & Bodaghi, 2012). Como consecuencia, aquellos usuarios que carecen del acceso hacia los datos financieros y de la experiencia suficiente para indagar acerca de la información financiera, tienden a tener menos conocimientos sobre el valor razonable de los instrumentos financieros (Evans, 2005).

Por otra parte, la administración corporativa de las empresas suele ser compleja y sujeta a muchas leyes y reglamentos. Recientemente, los reguladores, las organizaciones profesionales y las entidades normativas de información financiera en todo el mundo han mirado a XBRL y los datos interactivos como una forma de promover la transparencia de la información financiera y el seguimiento de la información empresarial (Roohani, Furusho, & Koizumi, 2009). En términos más amplios, XBRL es una tecnología mediante la cual se facilitan las búsquedas y la presentación simultanea de Estados financieros relacionados, los Estados financieros basados en XBRL proporcionan a los usuarios la oportunidad de buscar directamente la información que les resulte pertinente, independientemente de la ubicación en la que esta se encuentre y les permite comparar convenientemente la información relacionada entre diferentes empresas. Por consiguiente, XBRL es una alternativa de solución para reducir la falta de fiabilidad de la información financiera de las empresas ayudándoles a minimizar el efecto negativo de las decisiones que se basan en la información que se analiza a partir de los Estados financieros (Ahmadpour & Bodaghi, 2012).

Constituida como una empresa en Abril del año 2000, XBRL International[8] ha supervisado el desarrollo del estándar XBRL (Doolin & Troshani, 2004). Conforme ha pasado el tiempo, este estándar ha sido adoptado por diversas organizaciones alrededor del mundo, un ejemplo de ello es que desde Abril del año 2005 la SEC ha exhortado a los contribuyentes a proporcionar voluntariamente documentos siguiendo el estándar XBRL como ficheros adjuntos en el Sistema EDGAR *(EDGAR Company filings)* (Gray & Miller, 2009). El Sistema EDGAR fue desarrollado para ofrecer a los usuarios un medio eficiente para la preparación y el intercambio de informes de negocios, especialmente para la información financiera a través de Internet (Gerdes, 2003; Kambil & Ginsburg, 1998). Aunado a esto, varias bolsas de valores, autoridades fiscales y otros organismos reguladores

[8]XBRL International: http://www.xbrl.org/Jurisdictions

están exigiendo a las organizaciones que dependen de ellos el uso del estándar XBRL, tal uso parece ser fácilmente aceptado por los usuarios que son tolerantes a cuestiones relacionadas con la adopción de nuevas tecnologías y que están menos preocupados por un retorno a corto plazo de la inversión que requiera su implementación, así como por la mayor parte del mercado potencial, quienes son más sensibles a los costes, el retorno de la inversión y la facilidad de uso de XBRL (Gray & Miller, 2009).

2.4 Estado del arte de las tecnologías semánticas

Con la intención de organizar y complementar el estado del arte correspondiente a este trabajo de tesis, en esta sección se incluyen investigaciones e información teórica y técnica en relación con proyectos basados en tecnologías de la Web Semántica que permiten sustentar el proceso de transformación de datos XBRL para la creación de la base de conocimientos financieros que se describe en el Capítulo 4. Se parte con la descripción de la propia Web Semántica, las ontologías, así como los lenguajes y herramientas para su desarrollo, los modelos de datos y tecnologías relacionadas directamente con Linked Data (Linked Open Data).

2.4.1 Web Semántica

En los último años el crecimiento de la Web tradicional ha sido constante por ejemplo, para el año 2006 se calculó la existencia de 10 mil millones de páginas Web (Shadbolt, Hall, & Berners-Lee, 2006). La ingente cantidad de páginas Web disponibles a través de Internet y cuyo contenido está dirigido para la lectura, manipulación y compresión humana, permite suponer la presencia de diversas dificultades de extracción y organización de la información disponible en esas páginas por tratarse de información en un formato no estructurado. Por consiguiente, es natural pensar en la necesidad de una evolución en la Web tradicional, es decir, una evolución que proporcione soluciones a cada una de las dificultades previamente mencionadas y otros posibles inconvenientes. Esta evolución se encauzó hacia la Web Semántica, la cual surgió a través de la reestructuración y el enriquecimiento de las páginas y los componentes de la Web tradicional mediante la incorporación de información Semántica explicita destinada a ser comprendida por los ordenadores, independiente de la presentación al usuario y susceptible de ser procesada de forma automática por aplicaciones informáticas (T Berners-Lee, Hendler, & Lassila, 2001). Por lo tanto, la Web Semántica es una Web de nueva generación en la que los contenidos son más que un enorme cúmulo de información y servicios escasamente estructurados.

La Web Semántica no es una Web independiente, es una extensión de la Web tradicional cuyo éxito se fundamenta con la integración de un conjunto de tecnologías tal y como se ejemplifica en la Figura 2, entre las cuales la ontología es la principal (Castellanos-Nieves, Fernández-Breis, Valencia-García, Martínez-Béjar, & Iniesta-Moreno, 2011; Davies, Fensel, & Harmelen, 2003). Las ontologías definen puntos de vista comunes, compartibles y reutilizables, proporcionando sentido a las estructuras de información procesadas por los sistemas informáticos (Brewster & O'Hara, 2007). En su obra, (Gruber, 1993) complementa esta descripción definiendo la ontología como "*la especificación de una conceptualización*", lo que significa que una ontología es una descripción de conceptos y sus relaciones existentes para un agente o una comunidad de agentes cuyo propósito es compartir y reutilizar conocimiento.

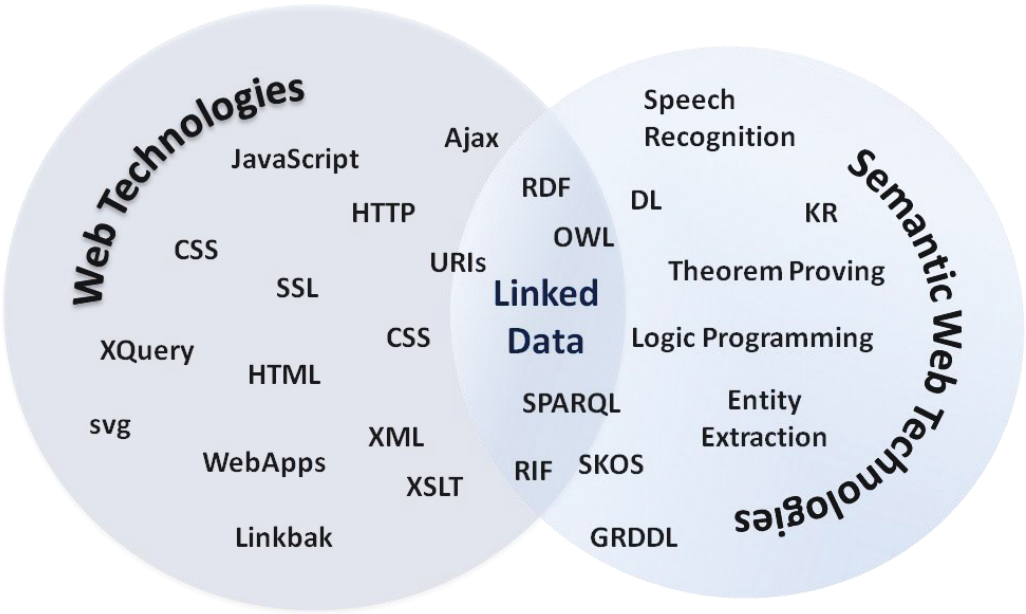


Figura 2. Tecnologías Web y tecnologías de la Web Semántica (Colomo-Palacios et al., 2012)

La integración de tecnologías Semánticas en la Web tradicional proporciona considerables mejoras en los resultados que se obtienen a través de la ejecución de aplicaciones informáticas aplicables en diversos dominios. Existen varios ejemplos de estas aplicaciones, uno de ellos son las Anotaciones Semánticas cuyo proceso permite añadir metadatos semánticos a los recursos Web y consiste en asignar significado inequívoco a este tipo de recursos con el fin de habilitar un mecanismo de descubrimiento de información más eficiente (Kiryakov et al., 2004). Otro ejemplo de gran utilidad es el descubrimiento semántico de contenido, este tipo de descubrimiento hace uso de taxonomías y ontologías para identificar y describir los diferentes tipos de contenido para la vinculación dinámica entre estos elementos (A. Sheth et al., 2002). La integración de

servicios es otra aplicación de las tecnologías semánticas y consiste en añadir semántica a los servicios Web con el objetivo de aumentar la integración de sus servicios, entre los que cabe destacar, el descubrimiento, la composición, la clasificación, la selección y la mediación de los servicios (Janev & Vraneš, 2011). Asimismo, las Interfaces de lenguaje natural resultan de gran importancia ya que se basan en la aplicación de métodos, técnicas y tecnologías útiles para proporcionar una interacción natural entre los humanos y los ordenadores (Kaufmann & Bernstein, 2010). Los Métodos de búsqueda semánticos tienen el objetivo de aumentar y mejorar los resultados de búsqueda tradicionales mediante el uso de no únicamente palabras, sino conceptos y las relaciones lógicas que permiten obtener una información más precisa. Un último ejemplo es el uso de tecnologías semánticas en las redes sociales, estas tecnologías se utilizan en las redes sociales para ayudar a las personas a rastrear, descubrir y compartir contenidos en torno a los temas que les interesan (Dietrich, Jones, & Wright, 2008).

Gracias a la integración de tecnologías de la Web Semántica, la Web tradicional ha pasado de ser un espacio de información global de documentos vinculados, a uno en el que se vinculan ambos, tanto documentos como datos. Uno de los beneficios de esta evolución es un conjunto de mejores prácticas para la publicación y la conexión de datos estructurados en la Web conocida como Linked Data (Segaran, Evans, & Taylor, 2009).

2.4.2 Ontologías

La definición de ontología proporcionada por Gruber, (1993), descrita en la sección 4.2 de esta tesis menciona que una ontología se compone de cuatro tipos de componentes: relaciones, funciones, axiomas e instancias. Estos componentes se describen brevemente a continuación:

- **Conceptos:** son las ideas básicas que tratan de ser formalizadas. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, entre otros.
- **Relaciones:** las relaciones representan la interacción y enlace entre los conceptos de dominio.
- **Funciones:** una función es un tipo especial de relación en la que un elemento se identifica mediante el cálculo de una función.
- **Axiomas:** son teoremas sobre las relaciones que deben cumplir los elementos de la ontología.

Por otra parte, (Borst, 1997) extiende tal definición mencionando que *"la especificación de una ontología explicita, es una conceptualización compartida"*. De la misma manera, (Studer, Benjamins, & Fensel, 1998) complementan esta definición señalando que una ontología es "*una especificación formal y explícita de una conceptualización compartida"*. En esta última definición, los autores mencionan que la conceptualización, es un modelo abstracto de algún acontecimiento en el mundo basado en la identificación de los conceptos de mayor relevancia de ése acontecimiento, que una ontología es explicita porque el tipo de conceptos que se utilizan y las limitaciones de su uso se definen claramente, también menciona que la formalidad de las ontologías radica en el hecho de que sean legibles para la máquina excluyendo el lenguaje natural. Finalmente, puntualizan que una conceptualización compartida refleja la noción de que la ontología captura conocimiento consensual, dicho en otros términos, se trata de conocimiento que no es privado de un individuo y que es aceptado por un grupo.

Existen varios tipos de ontologías orientadas a objetivos diferentes. Algunos ejemplos son descritos brevemente a continuación.

Las ontologías para la representación del conocimiento, permiten capturar las primitivas de representación necesarias para formalizar el conocimiento para un campo determinado (Gruber, 1995). Un ejemplo de esta ontología es *Frame-Ontology* (Gruber, 1993), que captura las primitivas de representación utilizadas en Lenguajes Basados en Marcos *(Frame-Based Languages)*. Por otra parte, las ontologías generales o comunes incluyen un vocabulario relacionado con las cosas, eventos, tiempo, espacio y comportamiento, entre otros (Mizoguchi & Ikeda, 1998). Un ejemplo de una ontología general es Cyc-Project (Lenat & Guha, 1989), que cuenta con una considerable cantidad de conocimiento fundamental humano. Las meta-ontologías son también llamadas *Ontologías genéricas o básicas* y permiten su uso en cualquier dominio independientemente de la naturaleza de esta. Un meta-ontología es un esquema para desarrollar ontologías que cumplan con determinados requisitos y que éstos cubran las propiedades importantes correspondientes al dominio para el cual ha sido destinada (Dietz & Habing, 2004).

2.4.3 Lenguajes para el desarrollo de ontologías

Una de las recomendaciones del consorcio de la W3C[9] *(World Wide Web Consortium)* para la representación de la información en la Web es RDF *(Resource Description Framework),*

[9]W3C: http://www.w3.org/

el cual fue desarrollado para la descripción de recursos Web y permite la especificación de la Semántica de los datos basados en XML de manera interoperable y estandarizada. Una de sus principales características es que proporciona mecanismos para representar explícitamente los servicios, procesos y modelos de negocio, al tiempo que permite el reconocimiento de la información no explícita (Cyganiak et al., 2014). El modelo de datos RDF equivale a una serie de formalismos aplicables a las redes estructuradas para la representación de conocimientos, las denominadas redes semánticas (Brachman, 1978), estas redes Semánticas que están compuestas por conjuntos de tripletas RDF también llamados grafos RDF (Cyganiak et al., 2014).

Para su funcionalidad, una tripleta RDF se compone de tres tipos de recursos: a) Sujeto *(Subject)*: siempre nombrados por URI's *(Uniform Resource Identifier,* Identificador Uniforme de Recursos*)*, además de los IDs de anclaje opcionales *(anchor IDs)*; b) Predicado o propiedad *(Predicate or property)*: define los aspectos específicos, características, atributos o relaciones utilizadas para describir un recurso además de expresar la relación existente entre el sujeto y el objeto; y c) Objeto *(Object)*: este asigna el valor de una propiedad a un recurso en específico (este valor puede ser otra declaración RDF) (DuCharme, 2011; Lassila & Swick, 1999). Un grafo RDF se visualiza como un nodo y un diagrama de arco dirigido, en el que cada tripleta es representada como un enlace Nodo–Arco–Nodo. La Figura 3, muestra los elementos de una tripleta basada en RDF.

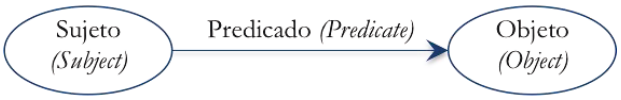


Figura 3. Elementos de una tripleta basada en RDF (Cyganiak et al., 2014).

En relación con RDF el W3C también recomienda RDF(S) *(RDFS Schema - Resource Description Framework Schema)* (Brickley & Guha, 2014) que a diferencia de RDF, RDF(S) proporciona funcionalidades para la definición de las relaciones entre las propiedades (atributos) y los recursos así como capacidades básicas para la descripción de vocabularios RDF incluyendo la posibilidad de añadir características adicionales que resultan de gran utilidad, tales características son proporcionadas a través de un mayor desarrollo en los RDF(S). La principal característica de RDF(S) es ofrecer primitivas para la definición de modelos de conocimiento que se acercan más a los enfoques basados en marcos. RDF(S) es ampliamente utilizado como un formato de representación en muchas herramientas y proyectos, como Amaya, Protégé, Mozilla y SilRI por mencionar algunos (Gómez-Pérez & Corcho, 2002).

La notación N3 *(Notation 3)*, proporciona una sintaxis basada en texto para RDF. Por lo tanto el modelo de datos de N3 está diseñado conforme al modelo de datos RDF. Además, N3 permite definir reglas, que se denotan mediante una sintaxis especial. Aunque tales reglas no definen a N3 como un lenguaje de consultas por sí mismo, sí permiten que sea utilizado con ese propósito. Para esto, es necesario que las consultas se almacenen como reglas en un fichero dedicado que se utilice en conjunción con los datos, de tal manera que el comando de filtro CWM *(Closed World Machine)* (Berners-Lee & Connolly, 2000) que permite seleccionar automáticamente los datos que se generan por las reglas, sea aprovechado. Y aunque N3 cumple con las propiedades de ortogonalidad, cierre y seguridad, resulta engorroso utilizarlo como lenguaje de consultas (Berners-Lee & Connolly, 1998). Respecto al manejo semántico de los datos, N3 está soportado por dos sistemas de acceso libre, Euler y CWM (Berners-Lee & Connolly, 2002) sin embargo, ninguno de estos dos sistemas se adhiere de manera automática a la semántica RDF, la semántica tiene que ser proporcionada por reglas personalizadas (Haase, Broekstra, Eberhart, & Volz, 2004).

El *Terse RDF Triple Language,* o *Turtle,* es una sintaxis textual para RDF que permite escribir grafos de este tipo, con un texto de forma compacta y natural, con las abreviaturas y tipos de datos de uso común (David Beckett & Berners-Lee, 2011). *Turtle,* ofrece niveles de compatibilidad con el formato N-Triples, así como la sintaxis del patrón de tripletas de SPARQL recomendado por el W3C (Prud'hommeaux, Carothers, & Machina, 2014). Además, *Turtle* permite describir nodos blancos *(Blank nodes),* también llamados recursos anónimos. Su propósito, es proporcionar la conectividad necesaria entre las distintas otras partes del grafo RDF sin la necesidad de utilizar URIs pero sí mediante la especificación de algún tipo de identificador explícito que permita representar a este como una tripleta en el grafo RDF. Los *Blank nodes* proporcionan una manera de hacer más precisa las declaraciones sobre los recursos que tienen URIs, pero que en términos de relaciones, se describen con otros recursos que sí tienen URIs (F Manola et al., 2007).

N-Triples, es un formato de texto plano para codificar grafos RDF. Originalmente, se definió como una sintaxis para los documentos de Casos de Prueba RDF. Debido a su popularidad como un formato para el almacenamiento e intercambio datos, el Grupo de Trabajo RDF decidió publicar una versión más actualizada (D Beckett, 2014). Este formato fue diseñado para ser un subconjunto fijo de N3 y por lo tanto, herramientas para N3 tales como CWM, Euler y N-TRIPLES2KIF (Yadagiri & Ramesh, 2013) son utilizadas para leer

y procesar documentos escritos en N-Triples. Para evitar confusiones entre N3 y N-Triples, se recomienda, (pero no se requiere), que el contenido de los documentos N-Triples se almacene en ficheros con el sufijo ". nt " (D Beckett & Barstow, 2001).

A principios de la década de los 90's, se creó un conjunto de lenguajes para la implementación de ontologías basadas en Inteligencia Artificial *(AI, Artificial Intelligence)*. Básicamente, el Paradigma para la Representación del Conocimiento *(KR Paradigm, Knowledge Representation paradigm)* que subyace a este tipo de lenguajes se basa en una Lógica de Primer Orden (KIF), en marcos combinados con la lógica de primer orden (Ontolingua, OCML y F-Logic) o DL *(Description Logic)* y (LOOM) (R. MacGregor & Bates, 1987).

Prosiguiendo con el contexto de la Web Semántica, en los párrafos siguientes se proporciona una breve descripción acerca de los lenguajes para el desarrollo de ontologías más relevantes.

El Formato para el Intercambio de Conocimientos o KIF *(Knowledge Interchange Format)* por sus siglas en inglés, es un lenguaje basado en la lógica de primer orden y creado como un formato de intercambio para los diferentes sistemas de representación del conocimiento. Es el más expresivo de los lenguajes que se utilizan para representar ontologías, lo que permite representar conceptos, taxonomías de conceptos, relaciones n-arias, funciones, axiomas, instancias y procedimientos. Sin embargo, el lenguaje por sí mismo no proporciona soporte para el razonamiento automatizado (Genesereth & Fikes, 1992; Ginsberg, 1991).

Ontolingua es un sistema para la descripción de ontologías en un formato canónico que le permite traducirlas fácilmente en una variedad de sistemas para la representación del conocimiento y razonamiento. Esto incluye la definición de clases, relaciones, funciones y objetos para mantener a la ontología en un formulario único y legible por la máquina mientras que es utilizada en sistemas con diferentes capacidades de sintaxis y de razonamiento. La sintaxis y la semántica de Ontolingua están basadas en KIF, lo que significa que Ontolingua extiende a KIF a través de primitivas estándar para la definición de clases y relaciones así como la organización del conocimiento en jerarquías centradas en el objeto con la herencia (Gruber, 1992, 1993).

LOOM (MacGregor & Bates, 1987) es un Lenguaje para la Representación del Conocimiento *(Knowledge Representation Language)* que ofrece una clasificación automática de conceptos y está basado en lógicas descriptivas y normas de producción. Además, permite

representar conceptos, taxonomías de conceptos, relaciones n-arias, funciones, axiomas y reglas de producción (MacGregor, 1991).

El lenguaje OCML *(Operational Conceptual Modelling Language)* sirve de apoyo para el modelado a nivel de conocimiento. En la práctica, este rol implica que OCML se centra en la lógica en lugar de las primitivas del nivel de aplicación. Por lo tanto, proporciona mecanismos para expresar ítems tales como relaciones, funciones, reglas, clases e instancias para la resolución de problemas, en lugar de utilizar matrices o tablas hash (Motta, 1998). En adición, incluye mecanismos para la definición de ontologías y métodos de resolución de problemas (Motta, 1999).

F-Logic *(Frame Logic)* es un marco de referencia base para los lenguajes de descripción del conocimiento y es utilizado para proporcionar una relación completa de este tipo de lenguajes sin perder la Semántica directa y su carácter descriptivo (Balaban, 1995). F-Logic combina marcos y la lógica de primer orden, permite representar conceptos, taxonomías de conceptos, relaciones binarias, funciones, axiomas y reglas deductivas. Adicionalmente, proporciona un mecanismo de inferencia que se puede utilizar para la verificación de las restricciones y la deducción de nueva información (Kifer, Lausen, & Wu, 1995).

El progreso de la Web Semántica ha favorecido el surgimiento de otros lenguajes para el desarrollo de ontologías tales lenguajes permiten el aprovechamiento de un número mayor de características ofrecidas por la propia Web Semántica. A continuación, se proporciona información referente a cada uno de estos lenguajes.

SHOE *(Simple HTML Ontology Extensions)* es un lenguaje desarrollado en la Universidad de Maryland fue desarrollado como una extensión de HTML. SHOE utiliza etiquetas que permiten la inclusión de ontologías en los documentos HTML. También combina marcos y reglas que representan conceptos, taxonomías de conceptos, relaciones n-arias, instancias y reglas de deducción que son utilizados por un motor de inferencia para la generación de conocimiento (Heflin, Hendler, & Luke, 1999).

Es importante mencionar que diversos lenguajes para el desarrollo de ontologías están basados en XML, porque es ampliamente aceptado como el lenguaje estándar para el intercambio de información en la Web. XML describe una clase de objetos de datos llamados documentos XML que describen parcialmente el comportamiento de los programas que el ordenador procesa. XML es un perfil de aplicación de forma restringida del SGML *(Standard Generalized Markup Language,* Estándar de Lenguaje de Marcado

Generalizado*)* (Goldfarb, 1990) por lo tanto, la construcción de los documentos XML es conforme a este tipo de documentos (Bray et al., 1998).

XML facilita datos eficientes porque el intercambio de datos codificados en este lenguaje es auto-descrito, de manera que los datos se intercambian y procesan sin modificación alguna. XML también ofrece búsquedas más significativas, ya que desde XML se proporciona información de contexto por medio de la codificación de información adicional en las etiquetas que describen el contenido y su estructura, lo que se traduce a que las búsquedas generen resultados más precisos y relevantes. Además, el usuario puede ver los datos de diversas maneras mediante el uso de información estructurada en datos XML. Por último, XML es un estándar abierto e independiente de la plataforma, es la creación de una forma universal para el formato y la presentación de datos (Efrim-Boritz & No, 2005; Harold, 2004).

Un ejemplo relacionado con la descripción proporcionada en los dos párrafos previos es el lenguaje XOL *(XML-Based Ontology Exchange Language)* cuyo objetivo es el intercambio de ontologías en dominios biomédicos (Karp, Chaudhri, & Thomere, 1999). XOL se centra en el intercambio de ontologías entre diferentes sistemas de bases de datos o bases de conocimiento, herramientas de desarrollo de ontologías o programas para su aplicación, respectivamente. XOL permite definir un subconjunto de interfaces de acuerdo con el protocolo para la Conectividad de Bases de Conocimiento Abierto *(Open Knowledge Base Connectivity)* bajo la sintaxis de XML lo que permite describir clases, slots y facetas, sin embargo, no contempla el uso de marcos. XOL es muy restrictivo proporciona un mecanismo de inferencia y favorece la representación de conceptos, taxonomías de conceptos y relaciones binarias (Schumacher, 2003).

El OML *(Ontology Markup Language)* es un lenguaje de ontologías desarrollado en la Universidad de Washington. Inicialmente, OML fue desarrollado como una serialización XML del lenguaje SHOE y se compone de diferentes capas que incrementan su expresividad. La Semántica del OML en especial la de los niveles más altos se basa en gran medida en grafos conceptuales (Antoniou, Franconi, & Van-Harmelen, 2005) . Hay cuatro niveles diferentes de OML, estos son: *OML Core* que se relaciona con los aspectos lógicos del lenguaje y se incluye en el resto de las capas; *Simple OML* este nivel se correlaciona directamente con RDF(S); *Abbreviated OML* incluye características gráficas conceptuales; y *Standard OML* que es la versión más expresiva de OML (Gómez-Pérez & Corcho, 2002).

OIL *(Ontology Interchange Language)* es un lenguaje que permite la interoperabilidad Semántica entre los recursos Web. Tanto su sintaxis como su Semántica se basan en los lenguajes OKBC, XOL y RDF(S). OIL proporciona primitivas de modelado de uso común dentro de los enfoques basados en marcos de la ingeniería ontológica, incluyendo los conceptos, las taxonomías de conceptos y las relaciones. El soporte a la Semántica formal y el razonamiento se encuentra en los enfoques de la descripción lógica. Tales enfoques son un subconjunto de la lógica de primer orden con un alto poder expresivo, funcionalidades de decidibilidad *(decidability)* y un mecanismo de inferencia. En este contexto, el término decidibilidad se refiere a identificar y dar solución a los problemas de decisión a tratar a través de un conjunto de fórmulas, teoremas o algoritmos (Krajewski, 1981). La relación entre OIL y RDF/RDF(S) es muy cercana ya que los tres están destinados a captar el significado en forma de redes Semánticas además, del mismo modo que RDF(S), OIL puede ser utilizado para definir otros lenguajes de ontologías (Fensel et al., 2000).

Relacionado a OIL se encuentra DAML+OIL *(DARPA, Agent Markup Language + OIL)*. Se trata de un lenguaje de marcado semántico para recursos Web, el cual se fundamenta en los estándares RDF y RDF(S) que son recomendados por el W3C. Estos lenguajes son extendidos por DAML+OIL mediante primitivas enriquecidas para el modelado, con una Semántica limpia y bien definida basada en lógicas descriptivas (Horrocks, Van-Harmelen, & Patel-Schneider, 2001). El objetivo de DAML+OIL es apoyar a la transformación de la Web para que pase de ser un foro para la presentación de información a un recurso que permita la interoperabilidad y el razonamiento de los datos (Mcguinness., 2002).

Un lenguaje derivado de DAML+OIL es OWL *(Web Ontology Language)* (Antoniou & Van-Harmelen, 2004). OWL es un lenguaje de marcado semántico para publicar y compartir ontologías en la Web, fue desarrollado como una extensión del vocabulario RDF y aunque se limita a el uso de reglas basadas en árboles, incluye funciones de conjunción, disyunción y manejo de variables existencial y universalmente cuantificadas. Adicionalmente, los sistemas de razonamiento se pueden beneficiar mediante el uso de OWL para realización inferencias lógicas y la obtención de conocimiento (Bechhofer, 2009). Sin embargo, su expresividad tiene algunos inconvenientes ya que la descripción y construcción de algunas ontologías pueden resultar ser muy complejas, motivo por el que OWL se clasifica en tres sub-lenguajes cada vez más expresivos y diseñados para el uso de comunidades específicas de implementadores de ontologías y usuarios interesados en el

diseño y desarrollo de ontologías (Berendt et al., 2004). Los sub-lenguajes de OWL se describen brevemente a continuación (McGuinness & Van-Harmelen, 2004):

- **OWL Full:** está destinado a los usuarios que requieren la máxima expresividad y la libertad sintáctica del RDF sin garantías computacionales. Por ejemplo, en OWL Full una clase puede ser tratada simultáneamente como una colección de individuos y como individuos por derecho propio. OWL Full proporciona una ontología para incrementar el significado del vocabulario predefinido ya sea RDF u OWL. Si consideramos todas las características que este sub-lenguaje de OWL ofrece, resulta poco probable que cualquier Software de razonamiento tenga las capacidades suficientes para darle soporte por completo.
- **OWL DL:** este sub-lenguaje es aprovechado por los usuarios que requieren de una máxima expresividad conservando completitud computacional y decidibilidad. Esto significa que se garantiza que todas las conclusiones son computables y que todos los cálculos terminarán en tiempo finito. OWL DL incluye todas las construcciones del lenguaje OWL, pero sólo puede ser utilizado bajo ciertas restricciones (por ejemplo, mientras que una clase puede ser una subclase de muchas clases, una clase no puede ser una instancia de otra clase). OWL DL es llamado así debido a su correspondencia con lógicas descriptivas, un campo de investigación que ha estudiado las lógicas que forman la base formal de OWL.
- **OWL Lite:** es útil para aquellos usuarios que requieren una jerarquía de clasificación y restricciones simples ya que se trata del un sub-lenguaje más restringido de OWL. Entre sus características se incluyen que sólo permite valores de cardinalidad de 0 o 1, tiene una complejidad formal menor que OWL DL y proporciona una ruta de migración rápida para tesauros y otras taxonomías.

OWL 2 *(Web Ontology Language 2)* es un lenguaje para el desarrollo de ontologías con significado formalmente definido para la Web Semántica. Las ontologías desarrolladas en OWL 2 proporcionan clases, propiedades, individuos y el valor de sus datos. Estas ontologías se almacenan como documentos propios de la Web Semántica y pueden ser utilizadas con información escrita en RDF ya que por sí mismas, tienen la función de intercambiarse en este tipo de documentos.

OWL 2 tiene una estructura general muy similar a OWL también nombrado OWL 1. Sin embargo, OWL 2 añade nuevas funcionalidades con respecto a OWL 1. Una de ellas es

el azúcar sintáctica *(Syntactic Sugar)*. (Un ejemplo de esto es la unión de la desunión de las clases). En este contexto, el término azúcar sintáctica se refiere a los agregados a la sintaxis de un lenguaje de programación que no afectan su funcionalidad y que facilitan la expresión de algunas de sus construcciones de manera más clara o concisa y posiblemente con un estilo alternativo. (Landin, 1964; Raymond, 1996). Otras funcionalidades que OWL 2 ofrece consisten en mejorar su expresividad mediante claves, cadenas de propiedades, tipos de datos enriquecidos, rangos de datos, restricciones de cardinalidad calificada, propiedades asimétricas, reflexivas y disjuntas (W3C-Group, 2012). Además de ofrecer capacidades de anotación mejoradas OWL 2 proporciona tres nuevos perfiles, estos perfiles son sub-lenguajes (subconjuntos sintácticos) que brindan importantes ventajas, particularmente en los escenarios de aplicación. Cada perfil es descrito brevemente a continuación (Grau et al., 2008):

- **OWL 2 EL:** se basa en la familia de lógicas descriptivas EL++, que han sido diseñadas para permitir el razonamiento eficiente con grandes terminologías. El interés principal de este razonamiento es computar la relación entre todas las clases y sub-clases contenidas en una ontología. El razonamiento en este perfil puede ser implementado en tiempo polinómico *(Polynomial Time)* basándose en el tamaño de la ontología. (Baader, Brandt, & Lutz, 2005). En términos de computación, el tiempo polinómico ayuda a resolver los problemas que se presentan cuando el tiempo de ejecución de un algoritmo es menor que el valor calculado a partir del número de variables implicadas en tal algoritmo, utilizando una fórmula polinómica. Las características centrales en el modelado de este perfil son la conjunción de clases y algunos valores de las restricciones. Con el fin de alcanzar tratabilidad, no está permitido el uso de la negación, la disyunción, todos los valores de las restricciones y las restricciones de cardinalidad. Muchas ontologías de gran escala pueden ser capturadas usando este perfil. En particular OWL 2 EL, captura un patrón muy común que se utiliza en las ontologías para definir conceptos, es decir, el uso combinado de la conjunción y la cuantificación existencial. Un ejemplo de esto, es afirmar que cada corazón contiene un ventrículo izquierdo y un ventrículo derecho.
- **OWL 2 QL:** este perfil se basa en la familia de lógicas descriptivas DL-Lite (Calvanese et al., 2007). Fue diseñado para permitir el razonamiento eficiente con grandes cantidades de datos estructurados de acuerdo con esquemas relativamente simples. Permite realizar consultas conjuntivas a través de tecnologías de bases de

datos relacionales estándar, y resulta muy conveniente para desarrollar aplicaciones donde se utilizan ontologías relativamente ligeras para la organización de un gran número de individuos en los que es necesario acceder directamente a los datos a través de consultas relacionales (por ejemplo, SQL, *Structured Query Language*). El perfil OWL 2 QL proporciona la mayoría de las características necesarias para capturar modelos conceptuales, tales como diagramas de clases UML *(Unified Modeling Language)*, diagramas Entidad Relación (ER) y esquemas de bases de datos.

- **OWL 2 RL:** este perfil ha sido diseñado de tal manera que varias tareas de razonamiento puedan ser implementadas como un conjunto de reglas en un sistema de reglas de encadenamiento progresivo. Para lograr la aplicabilidad de este criterio, OWL 2 RL no es tan expresivo como OWL 2, característica que lo hace atractivo en situaciones en las que se requiere de una extensión limitada de RDF(S). OWL 2 RL permite la mayoría de las construcciones de OWL 2, sin embargo, para permitir implementaciones basadas en reglas de razonamiento, la manera en la que estas construcciones pueden ser utilizadas en axiomas ha sido restringida. Estas restricciones aseguran que el motor de razonamiento sólo tiene que razonar con los individuos que se dan de forma explícita en la ontología. Además, OWL 2 RL también proporciona un conjunto de implicaciones de primer orden que se pueden aplicar directamente a un grafo RDF con el fin de derivar las consecuencias pertinentes. Estas implicaciones son una reminiscencia de la Semántica pD* de OWL 1 que proporcionan un punto de partida útil para la implementación de los razonadores de encadenamiento progresivo en OWL 2 RL. En este contexto, la Semántica pD* se refieren a una Semántica no estándar la cual se define de manera análoga a la Semántica de RDF(S) pero siendo más débil que la Semántica estándar de OWL 1 (ter Horst, 2005).

Para complementar la información proporcionada en esta sección, la Figura 4 muestra una pila que incluye a los lenguajes para el desarrollo de ontologías previamente descritos.

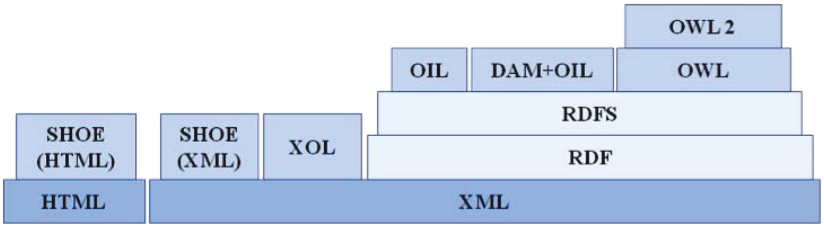


Figura 4. Pila de lenguajes para el desarrollo de ontologías. Figura basada en Corcho et al., (2003)

Con la evolución de la Web tradicional hacia la Web Semántica ha sido necesario el desarrollo de nuevos lenguajes de programación (y la mejora de los ya existentes) que proporcionen apoyo a esta nueva Web. En esta sección se presentó una recopilación sobre los lenguajes de programación para el desarrollo de ontologías propias de la Web Semántica, por lo que, a fin de complementar esta información, en la sección siguiente se presentan algunas de las herramientas para el desarrollo de aplicaciones dirigidas hacia la esta Web.

2.4.4 Herramientas para el desarrollo de aplicaciones de la Web Semántica

En los últimos años, el número de entornos y herramientas relacionadas con el desarrollo de ontologías ha mejorado considerablemente. Estas herramientas ayudan a proporcionar apoyo durante los procesos de desarrollo e implementación de las ontologías.

La primer herramienta para el desarrollo de ontologías a describir es *Ontolingua Server*, la cual fue desarrollada en el Laboratorio de Sistemas de Conocimiento (KSL, *Knowledge Systems Laboratory*)[10] de la Universidad de Stanford (Farquhar, Fikes, & Rice, 1997). Ontolongua Server se desarrolló a principios de 1990 y fue construido para facilitar el desarrollo de ontologías a través de una aplicación Web basada en formularios. Inicialmente, el principal módulo de Ontolingua Server fue el editor de ontologías, posteriormente, otros módulos fueron incluidos en su entorno, tales como el solucionador de ecuaciones Webster, el servidor OKBC *(Open Knowledge Based Connectivity)* y una herramienta para la fusión de ontologías llamada Chimaera (McGuinness et al., 2000). El editor de ontologías también proporciona un traductor para lenguajes como LOOM, PROLOG, IDL de CORBA y CLIPS *(C-Language Integrated Production System)*, por mencionar algunos. Además, los editores remotos pueden buscar y editar ontologías en él, y las aplicaciones remotas o locales pueden acceder a cualquiera de las ontologías disponibles en su biblioteca de ontologías mediante el protocolo OKBC (Chaudhri et al., 1998).

En el Instituto de Ciencias de la Información *(ISI, Information Sciences Institute)* de la Universidad del Sur de California desarrollaron la herramienta para la búsqueda y edición de ontologías OntoSaurus *(OntoSaurus Ontology Tool)* (Swartout et al., 1996a). Esta herramienta consta de dos módulos, el primero es un servidor de ontologías que emplea al lenguaje LOOM como su sistema para la representación del conocimiento, y el segundo es un browser Web para las ontologías desarrolladas en este lenguaje. Adicionalmente,

[10] KSL: http://ksl.stanford.edu/

OntoSaurus dispone de traductores del LOOM hacia los lenguajes siguientes Ontolingua, KIF, KRSS y C++. Con la herramienta OntoSaurus, las ontologías también se encuentran disponibles a través del protocolo OKBC (Swartout et al., 1996b).

La necesidad de representar y gestionar el conocimiento en diversos ámbitos a través de tecnologías Semánticas, conllevó a que diversas instituciones hayan creado herramientas para el desarrollo de ontologías cada vez más completas, un ejemplo de ello son Tadzebao y WebOnto desarrolladas en 1997 por el Instituto de Medios del Conocimiento *(KMI, Knowledge Media Institute)*[11] de la *Open University* en el Reino Unido. Tadzebao permite a los ingenieros del conocimiento mantener conversaciones síncronas y asíncronas sobre las ontologías abordando el hecho de que en la parte integral del diseño comunitarios, el dialogo ha sido ignorado por la comunidad. Por otra parte, WebOnto es un editor de ontologías basadas en OCML. Su ventaja principal es que da soporte a la edición colaborativa de ontologías permitiendo discusiones síncronas y asíncronas sobre las ontologías que están en desarrollo (Domingue, 1998).

La principal similitud entre las herramientas para el desarrollo de ontologías descritas previamente, consiste en que todas ellas tienen una fuerte relación con un lenguaje en específico (Ontolingua, LOOM y OCML, respectivamente). En realidad, estas herramientas fueron creadas para permitir la fácil navegación y edición de ontologías basadas en esos lenguajes, están orientadas hacia actividades de investigación y la mayoría de ellas fueron construidas como herramientas aisladas que no ofrecían muchas facilidades de extensibilidad. Hoy en día, existen nuevas herramientas para desarrollar ontologías que ofrecen funcionalidades más completas entre las que destacan la integración de la tecnología de las ontologías en los sistemas de información actuales, el uso de *plugins* y arquitecturas basadas en componentes. Además, los modelos de conocimiento que subyacen a estos entornos son independientes del lenguaje que se utilice para el desarrollo de las ontologías, entre estos entornos merece la pena describir Protégé, WebODE y OntoEdit (Corcho et al., 2003).

Protégé es una herramienta *Standalone* de código abierto con una arquitectura extensible (Noy, Fergerson, & Musen, 2000). El primero en desarrollar esta meta-herramienta fue Mark Musen en 1987, la versión original era una pequeña aplicación, destinada a la construcción de herramientas para la adquisición de conocimiento por parte de algunos programas especializados en planificación médica. A partir de esta herramienta inicial, el

[11]KMI: http://kmi.open.ac.uk/

sistema Protégé se ha convertido en una plataforma extensible y duradera para el desarrollo de sistemas basados en conocimiento e investigación. En su versión actual, además de conservar como núcleo su editor de ontologías, Protégé se ejecuta en una variedad de plataformas, soporta extensiones de interfaz de usuario personalizadas, incorpora el protocolo de conectividad OKBC, mantiene una biblioteca de *plugins* que le añaden más funcionalidad además de permitir la exportación de ontologías hacia otros lenguajes incluyendo a F-Logic, OIL y XML, así como hacia los motores de inferencia Jess y Prolog. Adicionalmente, interactúa con los formatos de almacenamiento estándar tales como bases de datos relacionales, XML y RDF, y ha sido usado por cientos de individuos y grupos de investigación (Gennari et al., 2003).

Por otra parte, la herramienta WebODE desarrollada en el Laboratorio de Inteligencia Artificial de la Universidad Politécnica de Madrid (UPM), consiste en una suite basada en una arquitectura flexible para la ingeniería de ontologías (Arpírez et al., 2001). WebODE no se utiliza como una aplicación *Standalone*, sino como un servidor de aplicaciones con su propia interfaz Web. El núcleo de este entorno es un servicio de acceso a la ontología, que es utilizado por todos los servicios y aplicaciones conectados a su servidor de aplicaciones, principalmente por el editor de ontologías integrado en el WebODE que a su vez, ofrece varios servicios como la importación y exportación de ontologías hacia los lenguajes XML, RDF(S), OIL, DAML+OIL, F-Logic, Jess y Prolog, la edición de axiomas mediante el Constructor de Axiomas WAB *(WebODE, Axiom Builder),* la documentación y evaluación de ontologías así como la fusión de las mismas (Arpírez et al., 2003; Corcho et al., 2003).

Una de las herramientas actuales y previamente mencionadas pero no descritas para el desarrollo de ontologías es OntoEdit. Se trata de una herramienta similar a Protégé y WebODE, fue desarrollada por el AIFB *(Institute of Applied Informatics and Formal Description Methods)* en Alemania (Sure et al., 2002). OntoEdit es una herramienta flexible ya que su arquitectura se basa en *plugins* que le proporcionan la funcionalidad para navegar y editar ontologías. Entre los *plugins* más importantes, se encuentra el encargado de realizar inferencia utilizando un conjunto de lenguajes y herramientas llamado Ontobroker (Decker et al., 1999; Fensel et al., 1998), el cual consiste en mejorar el acceso a las consultas y los servicios de inferencia de la Web (Sure, Angele, & Staab, 2002).

Apache Jena es un marco de trabajo *(Apache Jena Framework)* que soporta la mayoría de los estándares RDF y actualmente es el marco de trabajo para el desarrollo de aplicaciones de la Web Semántica *(Semantic Web Framework)* más completo. Una de sus principales

características es dar soporte a diferentes razonadores que infieren conocimiento adicional entre los que se incluye el razonador OWL Pellet (Sirin et al., 2007). Apache Jena es un marco de trabajo basado en Java, fue desarrollado en los laboratorios de HP *(Hewlett-Packard)* hasta Octubre de 2009. Desde entonces, Apache Jena es desarrollado y apoyado por una comunidad de código abierto. Además de ofrecer soporte a RDF, RDF(S), OWL y SPARQL, Apache Jena ofrece la posibilidad de leer y escribir anotaciones RDF comunes tales como RDF/XML, N3 y N-Triples, incluyendo el uso de APIs como RedLand, que consiste en un conjunto de bibliotecas de Software libre desarrolladas en lenguaje C que proporcionan apoyo para RDF (D Beckett, 2002). Adicionalmente, con el proyecto Apache Jena se desarrolló el servidor RDF Joseki (y su antecesor Fuseki), que proporciona una interfaz bajo el protocolo HTTP para RDF y ofrece soporte a SPARQL (Lindörfer, 2010). Complementando la información de Apache Jena, cabe mencionar que las nuevas recomendaciones de la Web Semántica para RDF, RDF(S) y OWL, tienen como núcleo los grafos RDF. En este sentido, Apache Jena 2 es un conjunto de herramientas de segunda generación que se centra en los grafos RDF *(RDF Toolkit)* que proporciona APIs más enriquecidas que incluyen el soporte para otros aspectos de las recomendaciones RDF como contenedores y reificación. La API de Ontologías incluye soporte para RDF(S) y OWL incluyendo soporte avanzado para OWL Full (Carroll et al., 2004). Apache Jena 2 incluye la referencia RDF/XML como parser de facto proporcionando una salida del mismo tipo mediante una gama enriquecida de la gramática RDF/XML e incluyendo la compatibilidad para la entrada y salida de notaciones N3. Por último, los grafos RDF se almacenan en memoria o en bases de datos y tanto el lenguaje de consulta de Apache Jena RDQL *(RDF Data Query Language)* y la API Web son ofrecidas para el desarrollo de aplicaciones de la Web Semántica (Wilkinson et al., 2003).

Varias de las herramientas para el desarrollo de ontologías permiten el uso de reificación *(Reification),* en el contexto de las ciencias computacionales y particularmente de la Web Semántica, la reificación es un medio de proporcionar metadatos a los datos RDF (Alexander & Ravada, 2006). Esto significa que la reificación, permite a las tripletas adjuntarse como propiedades en otras tripletas (Powers, 2003) .

Dicha definición se complementa citando a Manola, Miller, & McBride, (2004), que menciona que la reificación de una tripleta se expresa utilizando el vocabulario incorporado en los datos RDF y permite tratar como un recurso a toda una tripleta, por lo que las

aserciones necesarias, se aplicarán sobre esa tripleta o recurso. La reificación, por lo tanto, permite a las tripletas que se adjunten como propiedades a otras tripletas.

Para concluir esta sección, es importante mencionar que el desarrollo de ontologías es un proceso orientado a la representación del conocimiento de diversos y complejos dominios por lo que en gran medida se requiere aprovechar los beneficios que las herramientas de apoyo proporcionan en este ámbito (Khondoker & Mueller, 2010). En los últimos años, los investigadores han desarrollado una gran cantidad de herramientas para el desarrollo de ontologías, como las mencionadas anteriormente. Sin embargo, el desarrollo de herramientas para desarrollar ontologías con lenguajes como DAML+OIL y RDF(S) continua en crecimiento, algunos ejemplos son OILed (Bechhofer et al., 2001), SWOOP (Kalyanpur et al., 2005), COE Tool (Hayes et al., 2005) y OntoTrack (Liebig & Noppens, 2004) por mencionar algunos.

2.4.5 Modelo Conceptual

El término Modelo Conceptual, se utiliza en varios dominios, incluyendo la ingeniería del conocimiento y la ciencia. Un Modelo Conceptual, es una representación abstracta de algo generalizado en casos particulares (Borah, 2002). Pace, (2000) menciona que un Modelo Conceptual implica la construcción de representaciones del conocimiento humano. Complementariamente, Robinson, (2006) proporciona una definición más amplia de Modelo Conceptual, en ella indica que se trata de una descripción *Non-Software* específica del modelo de simulación que se va a desarrollar, describiendo sus objetivos, sus entradas y salidas, el contenido, los supuestos y las simplificaciones del modelo. Adicionalmente, Robinson menciona que un Modelo Conceptual tiene las siguientes propiedades:

- La actividad del modelado conceptual es iterativa y repetitiva a través de todo el ciclo de su desarrollo.
- El Modelo Conceptual es una representación simplificada del sistema real.
- El Modelo Conceptual es independiente del código del modelo o del Software.
- Las perspectivas tanto de los usuarios como de los desarrolladores, son tomadas en consideración.

Cada unas de las definiciones de Modelo Conceptual proporcionadas en esta sección, sirve de preámbulo teórico para la descripción del modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009) que se proporciona en el Capítulo 4 de

este trabajo de tesis. Sin embargo, estas definiciones se complementa mediante conceptos y definiciones relacionadas con los Modelos de datos y los Modelos de datos Semánticos, los cuales se describen en las siguientes dos secciones.

2.4.6 Modelos de datos

Un modelo de datos es una colección de conceptos bien definidos matemáticamente que ayudan a expresar las propiedades estáticas y dinámicas de una aplicación con un uso de datos intensivo (Brodie, 1984). En relación con esta definición, la premisa de (Silberschatz, Sudarshan, & Korth, 2002), menciona que bajo la estructura de las bases de datos se encuentra el modelo de datos, y que este consiste en una colección de elementos conceptuales que describen los datos, las relaciones, la semántica y las restricciones de consistencia. Siguiendo este contexto, los mismos autores resaltan que de acuerdo con su nivel de abstracción, los diferentes modelos de datos se clasifican en tres grupos diferentes: a) Modelos lógicos basados en objetos: se trata de modelos que están orientados a la descripción de estructuras de datos y restricciones de integridad; b) Modelos lógicos basados en registros: son modelos que no se orientan hacia la descripción de una realidad, sino que son orientados a las operaciones y cuya singularidad es poseer buenas características conceptuales como la normalización de datos; y c) Modelos físicos: son estructuras de datos a bajo nivel implementadas dentro del propio Sistema de Gestión de Base de Datos *(DBMS, DataBase Management System)*. Dos de los modelos de datos más ampliamente adoptados son el modelo entidad-relación, que forma parte del grupo de los modelos lógicos basados en objetos (Chen, 1976) y el Modelo relacional que corresponde al grupo de los modelos basados en registros (Codd, 1970). Ambos modelos son descritos brevemente en las dos secciones siguientes.

2.4.6.1 Modelo de datos entidad-relación

El modelo de datos entidad-relación comúnmente nombrado modelo E-R *(E-R Model, Entity–Relationship Model)* está basado en una percepción del mundo real que consta de una colección de objetos básicos, llamados entidades y de relaciones entre estos objetos (Chen, 1976). Una entidad es una *cosa* u *objeto* en el mundo real que es distinguible de otros objetos. Las entidades se describen en una base de datos mediante un conjunto de atributos de los cuales, es una buena práctica de modelado el uso de un atributo extra que sirva de identificador único con el objetivo de evitar la inconsistencia de datos. Finalmente, una relación es una asociación entre varias entidades, el conjunto de todas las entidades del

mismo tipo, y el conjunto de todas las relaciones del mismo tipo, se denominan respectivamente conjunto de entidades y conjunto de relaciones (Teorey, 1990).

2.4.6.2 Modelo relacional

En el Modelo relacional *(Relational Model)* se utiliza un grupo de tablas para representar los datos y las relaciones entre ellos. Cada tabla está compuesta por varias columnas, y cada columna tiene un nombre único. Este modelo es un ejemplo de un modelo basado en registros, los modelos basados en registros se denominan así porque la base de datos se estructura en registros de formato fijo de varios tipos. Cada tabla contiene registros de un tipo particular. Cada tipo de registro de fine un número fijo de campos, o atributos. Las columnas de la tabla corresponden a los atributos del tipo de registro (Codd, 1970; Date & Faudón, 2001). El Modelo relacional es el modelo de datos más ampliamente usado, y una extensa mayoría de sistemas de bases de datos actuales están basados en este modelo, su nivel de abstracción es inferior al modelo E-R por lo que los diseños de bases de datos a menudo se realizan en el modelo E-R, y posteriormente se traducen al Modelo relacional (Silberschatz et al., 2002).

Históricamente, otros dos modelos de datos precedieron al Modelo relacional, el modelo de datos de red (Bachman et al., 1969; Taylor & Frank, 1976) y el modelo de datos jerárquico (Tsichritzis & Lochovsky, 1976). Estos modelos estuvieron ligados fuertemente a la implementación subyacente y complicaban la tarea del modelado de datos (Silberschatz et al., 2002).

2.4.6.3 Otros modelos de datos

Existen otros modelos de datos más recientes, uno de ellos es el modelo de datos orientado a objetos *(OODM, Object-Oriented Data Model)* el cual es considerado como una extensión del modelo E-R con las nociones de encapsulación, métodos (funciones) e identidad de objetos (Norrie, 1994; Worboys et al., 1990). Otro modelo es el Modelo de Datos Relacional Orientado a Objetos *(Object-Relational Modeling)* que combina las características del modelo de datos orientado a objetos y el modelo relacional (Rumbaugh et al., 1991). Además de estos modelos, los modelos de datos semiestructurados permiten la especificación de datos donde los elementos de datos individuales del mismo tipo tienen diferentes conjuntos de atributos. Esto es diferente de los modelos de datos mencionados anteriormente, en los que cada elemento de datos de un tipo particular tiene el mismo conjunto de atributos. El lenguaje XML *(XML, eXtensible Markup Language)* se usa

ampliamente para representar datos en un formato semiestructurado (Abiteboul, Buneman, & Suciu, 2000).

2.4.7 Modelos de datos Semánticos

Los modelos semánticos se introdujeron en la década de los 70's (Abrial, 1974). Desde entonces, la investigación ha dado como resultado el desarrollo de poderosos sistemas para la representación de aspectos estructurales de los datos así como de sus características dinámicas y de comportamiento. Inicialmente, los modelos semánticos fueron desarrollados para facilitar el diseño de esquemas de bases de datos (Chen, 1976; Hammer & Mc Leod, 1981; Schmid & Swenson, 1975; Smith & Smith, 1977), un esquema era expresado utilizando las abstracciones de nivel superior de un modelo semántico, posteriormente, era traducido en alguno de los modelos tradicionales de bases de datos. Dado que el énfasis de los modelos semánticos era capturar las relaciones entre los datos, tal y como existen en la configuración del mundo real, estos tendían a facilitar las vistas que ofrecían la navegación entre los datos (Schnase et al., 1993).

Los sistemas de bases de datos relacionales expresan una comprensión muy limitada del significado de la información contendida en las bases de datos ya que por los general, sólo entienden ciertos valores y ciertas relaciones entre los datos almacenados en las tablas de la base de datos, a diferencia de estos modelos orientados a los registros que típicamente tienen dos o tres relaciones, los modelos de datos semánticos tienen la capacidad de expresar interrelaciones complejas de datos a través de construcciones que permiten especificar este tipo de relaciones (Hull & King, 1987). A pesar de las limitaciones que las bases de datos relacionales presentan para la expresión del significado de sus datos, son susceptibles de ser mapeadas hacia un modelo semántico (Christian Bizer, 2003). Con base en esta premisa, Berners-Lee, (2013) menciona que el modelo de datos de la Web Semántica se conecta muy directamente con el modelo de bases de datos relacionales lo que significa que el mapeo de las bases de datos relacionales hacia RDF es muy directo, como se menciona en la lista siguiente:

- Un registro es un nodo RDF
- El nombre del campo (Columna) es el RDF propertyType
- El campo del registro (Celda de la tabla) es un valor

Una de las principales fuerzas impulsoras de la Web Semántica siempre ha sido la expresión de la información, el hecho de que en la Web tradicional la gran cantidad de información que se manipula provenga de bases de datos relacionales, indica que tal información es aprovechable para ser procesada por las máquinas (Christian Bizer & Cyganiak, 2006) puesto que el formato de serialización de RDF (su sintaxis en XML) es muy adecuado para expresar la información de las bases de datos relacional (Berners-Lee, 2013).

Respecto al modelo E-R, Berners-Lee, (2013) menciona que básicamente el modelo RDF es una abertura del modelo E-R para trabajar en la Web. El modelo E-R típico involucra tipos de entidad y para cada tipo de entidad, hay un conjunto de relaciones (Llamadas *Slots* en este modelo). En este contexto, el modelo RDF es el mismo, a excepción de que las relaciones son objetos de primera clase que son identificados por una URI y que el conjunto de relaciones *(Slots)* de un objeto no se especifica cuando se define la clase de ese objeto. Por otro lado, también describe que la Web funciona a pesar de que (técnicamente) nadie esté autorizado a expresar algo. Esto significa que una relación entre dos objetos es capaz de almacenarse aparte de cualquier otra información acerca de los dos objetos. Esto es diferente de los sistemas orientados a objetos utilizados frecuentemente para implementar modelos E-R, que generalmente asumen que la información acerca de un objeto, es almacenada en un objeto, ya que la definición de la clase de un objeto define el almacenamiento implícito por sus propiedades.

Los componentes fundamentales utilizados por los modelos semánticos para estructurar los datos son: objetos, tipos atómicos y construidos, atributos, relaciones IS-A, y los componentes del esquema derivados. Cada uno de estos componentes es descrito brevemente a continuación (Schnase et al., 1993):

- **Objetos:** se asume que estos objetos son datos no estructurados o semiestructurados, tales como cadenas de texto, números enteros y reales, o, en el caso de recursos multimedia, voz e imagen (Christodoulakis, Ho, & Theodoridou, 1986; Woelk, Kim, & Luther, 1990).
- **Tipos atómicos:** La representación directa de los tipos de objetos distintos de sus atributos, es esencial para el modelado semántico. Como su nombre lo indica, los tipos atómicos corresponden a clases de objetos simples.
- **Tipos construidos:** los modelo semánticos se caracterizan por su capacidad para construir tipos de objetos complejos a partir de tipos atómicos. La agregación y

agrupación, también llamada asociación, son los constructores de tipos más comunes en la literatura semántica.

- **Atributos:** otro aspecto fundamental de los modelos semánticos es su capacidad para representar a las dependencias o conexiones interrelaciónales entre los tipos de objetos. Estas propiedades se denominan atributos o relaciones.
- **Relaciones IS-A:** prácticamente todos los modelos semánticos tienen la capacidad de representar relaciones IS-A o súpertipo/subtipo. En términos generales, una relación de IS-A a partir de un subtipo de un súpertipo indica que cada objeto asociado con el subtipo también se asocia con el súpertipo.
- **Componentes derivados del esquema:** también son llamados datos derivados y son un mecanismo básico para la abstracción y la encapsulación de datos en muchos modelos semánticos. La derivación permite que la información a incorporar en un esquema de base de datos, se calcule a sí misma a partir de otra información disponible en el esquema.

En resumen, los modelos de datos semánticos tienen el objetivo de capturar más significado para los datos mediante la integración de conceptos relacionales, dentro del campo de Inteligencia Artificial, este tipo de conceptos tienen mayor poder de abstracción. La idea consiste en proporcionar primitivas de modelado de alto nivel como parte integral de un modelo de datos con el fin de facilitar la representación de situaciones del *mundo real* (Klas & Schrefl, 1995).

2.4.7.1 Modelo de datos Entidad-Atributo-Valor

El modelo Entidad-Atributo-Valor *(EAV, Entity–Attribute–Value model)*, es popular para el modelado de datos altamente heterogéneos usando un relativamente simple esquema de base de datos físico (En la literatura de base de datos, los términos alternativos para las entidades y los atributos son los objetos y los parámetros, respectivamente). Un diseño EAV, conceptualmente implica una tabla con tres columnas, la primer columna es para la identificación de la entidad/objeto (ID), la segunda es para el atributo/parámetro (o el ID de un atributo que apunta a una tabla de descripciones de atributos), y la última es para el valor del atributo (Nadkarni et al., 1999). De acuerdo con (Anhøj, 2003), el diseño EAV tiene las siguientes ventajas:

- **Flexibilidad:** no hay límites arbitrarios en el número de atributos por entidad. El número de parámetros crecerá a medida que evoluciona la base de datos, sin la necesidad de rediseñar el esquema.
- **Eficiente espacio de almacenamiento de datos:** mientras que un diseño convencional conduce hacia campos vacíos (NULL) que requieren reservar espacio de almacenamiento, el diseño EAV no necesita reservar espacio para los atributos con valores NULL.
- **Consultas eficientes centradas en la entidad:** en una base de datos convencional si se necesita obtener toda la información para una sola entidad, será necesaria la consulta de todas las tablas de datos en busca de la información relacionada con esa entidad. Esto es una tarea que consume tiempo ya que requiere buscar a través de numerosas tablas, donde cada una puede o no tener información relacionada con la entidad buscada. Por lo contrario, en una base de datos EAV sólo es necesario consultar una tabla, no es necesaria la unión *(join)* con otras tablas, y no se requiere de ningún cambio de código a medida que evoluciona el dominio (una cláusula *join* combina datos de dos o más tablas basándose en un atributo común).

El modelo EAV se ha empleado en diversos sistemas por ejemplo, se utilizó por primera vez en aplicaciones de Inteligencia Artificial, mediante listas de asociación desarrolladas en lenguaje LISP *(LISt Processing Language)* (Bobrow & Murphy, 1967). Además, la estructura del modelo EAV es la base de las cookies Web, del registro de Microsoft Windows y de varios formatos para el intercambio de datos con etiquetas como ASN.1*(Abstract Syntax Notation One)* (Steedman, 1993). El lenguaje XML se considera una forma de etiquetado EAV (con etiquetas de apertura y cierre de atributos), que da soporte a la anidación de atributos en un grado arbitrario (Nadkarni, 1999).

2.4.7.2 Modelo de datos común

De acuerdo con (Dell & Dell, 2010) un modelo de datos común *(CDM, Common Data Model)* también nombrado modelo de datos canónico, define a las entidades pertinentes para un dominio en específico, incluyendo sus atributos, sus asociaciones y su semántica. El núcleo del CDM es la identificación de entidades abstractas a partir de las entidades que se derivan de los componentes de modelos de datos compatibles (Gardner et al., 2001), el modelo de referencia del CDM se basa en los siguientes seis principios básicos (Dell & Dell, 2010):

1. **Comprensibilidad:** refleja un vocabulario común del dominio. CDM no introduce abstracciones obtusas y está organizado de forma modular en vistas de fácil comprensión.
2. **Independencia:** el CDM es independiente de cualquier aplicación específica, pero sintetiza las necesidades de integración de datos de todas las aplicaciones relevantes.
3. **Kernel Inmutable:** las estructuras básicas (entidades y relaciones) del CDM son lógicamente comunes a través de todas las aplicaciones.
4. **Extensibilidad:** el CDM está diseñado para ser extendido para instalaciones particulares y futuros cambios y evoluciones.
5. **Mapeo conmutativo:** el proceso de mapear la información de modelos específicos en CDM preserva los datos y operaciones.
6. **Separación de asuntos *(Concerns)*:** el CDM separa lógicamente los datos de quién controla o es propietario de los datos. Por ejemplo, separar la información específica de una ciudad a través de una extensión en el modelo.

Hay dos razones para la definición de esquemas de componentes basados en CDM, la primera es que permite describir los esquemas locales divergentes mediante el uso de una sola representación y la segunda consiste en que si un sistema local carece de semántica, esta puede ser agregada a su esquema de componentes (Sheth & Larson, 1990).

Es importante mencionar que aunque la necesidad de hacer uso de modelos de datos con una semántica más enriquecida es ampliamente reconocida, no existe un enfoque único que haya ganado una aceptación general (Peckham & Maryanski, 1988).

2.4.8 Linked Data

Los datos enlazados (generalmente capitalizados como *Linked Data* en inglés), consiste en utilizar la Web para establecer vínculos entre los datos con tipo de contenido a partir de diferentes fuentes de datos. Estas fuentes de datos son tan diversas como: bases de datos mantenidas por varias organizaciones en diferentes ubicaciones geográficas, sistemas heterogéneos, o simplemente información existente dentro de una organización que históricamente no interopera fácilmente entre sí a nivel de datos. Técnicamente, Linked Data se refiere a la publicación de datos en la Web de tal manera que sean legibles por el ordenador, su significado se defina explícitamente, que estén relacionados con conjuntos de datos externos y que a su vez permitan ser vinculados desde conjuntos de datos externos (Christian Bizer, Heath, & Berners-Lee, 2009). Adicionalmente, Carroll & Klyne, (2004)

mencionan que la Web del hipertexto está conformada por páginas HTML conectadas por medio de hipervínculos sin tipo. A diferencia de estos, Linked Data de basa en documentos que contienen datos en formato RDF. Linked Data no sólo incluye la conexión entre los datos especificados en este tipo de documentos, también utiliza RDF para hacer declaraciones escritas que unen elementos de la Web tradicional con otras tecnologías de la Web semántica, tales como: SPARQL (DuCharme, 2011) y URI's (Masinter, Berners-Lee, & Fielding, 2005) entre otros, para mejorar su funcionalidad. Para la publicación de datos en la Web, (T Berners-Lee, 2009) describe un conjunto de reglas llamadas *Principios de Linked Data* que proporcionan una receta básica para la publicación y conexión de datos utilizando la infraestructura de la Web de forma que todos los datos publicados se convierten en parte de un único espacio global de datos:

- Utilizar las URI's para identificar los recursos publicados en la Web.
- Aprovechar el HTTP de las URI's para que los usuarios localicen y consulten (desreferencien) esos recursos.
- Proporcionar información útil acerca del recurso cuando la URI haya sido desreferenciada utilizando los estándares RDF y SPARQL.
- Incluir enlaces a otras URI's relacionadas con los datos contenidos en el recurso, por lo que es posible el descubrimiento de más información en la Web.

Linked Data se basa en dos tecnologías que son primordiales para la Web, las URI's (Masinter et al., 2005) y el protocolo HTTP (Fielding et al., 1999). Mientras que los Localizadores Uniformes de Recursos *(URL, Uniform Resource Locator)* son tratados como direcciones de documentos y otras entidades que son localizados en la Web, las URI's proporcionan un medio más genéricos para identificar de forma única a cualquier entidad en la Web. Por ende, la aplicación de las tecnologías de Linked Data junto con el correcto uso de su arquitectura y estándares proporciona un entorno que permite realizar varias operaciones entre las que se incluyen la consulta y la navegación entre los datos, la realización de procesos de inferencia así como la obtención de conclusiones utilizando vocabularios con información específica (Colomo-Palacios et al., 2012).

2.4.8.1 La importancia de Linked Data

Hoy en día la Web tradicional es una parte esencial en la vida de las personas ya que es utilizada para navegar de página en página con el objetivo de realizar múltiples actividades incluyendo la búsqueda de información, trabajar, estudiar e investigar, comunicarse, realizar

compras, divertirse, consumir y compartir contenidos, entre otros. Pese a que se trata de una Web tan completa y diversa, la idea central detrás de ella es realmente simple ya que consiste en unir documentos a través de Internet independientemente del ordenador en el que se encuentren albergados creando así una red de documentos (Jacobs & Walsh, 2004). De esta manera, los usuarios navegan de documento en documento utilizando enlaces lo que permite deducir que esta red crece en función de los documentos que se van creando y enlazando (Albert, Jeong, & Barabasi, 1999). Para hacer la Web posible, fue necesario el desarrollo de dos elementos técnicos básicos, el primero consiste en la utilización de un formato abierto y estándar para la representación de la información en los documentos de la red, es decir, las páginas Web. Tal estándar es el lenguaje HTML y el echo de que se trate de un formato abierto permite que los usuarios desarrollen aplicaciones de Software que consuma esos documentos, un ejemplo de ello son los navegadores Web. Por otra parte, el segundo elemento consistió en establecer un mecanismo estándar para localizar esos documentos en la red, tal elemento se refiere a los denominados URL's por lo que el lenguaje HTML incluye una manera de apuntar hacia ellos, es lo que conocemos como enlaces (Jacobs & Walsh, 2004). Sobre estas bases se asienta la Web, un espacio de información en el que los usuarios crean documentos HTML y los enlazan con otros documentos apuntando a sus URL's sin la necesidad de un sistema centralizado (Berners-Lee, 1992). De esta manera, la Web ha crecido exponencialmente desde su creación, siendo así una vastísima red de documentos en la que si los usuarios requieren buscar alguna información determinada, es necesario procesar los documentos que contienen esa información para llegar a ella. Sin embargo, esto representa un problema, ya que no permite buscar la información directamente en la Web. Es en este punto, en el que radica la importancia de Linked Data y en donde Berners-Lee, (2006) propone una manera más eficiente de codificar la información para solucionar las limitaciones de la Web tradicional. Esta solución consiste en utilizar la infraestructura Web para codificar y enlazar directamente la información en vez de hacerlo mediante documentos, el principio básico de Linked Data es dar un localizador a cada concepto que deseemos representar y enlazar los conceptos directamente en vez de enlazar a los documentos que contienen a esos conceptos, de este modo se obtiene una red de conceptos en lugar de una red de documentos que es lo que existe en la Web tradicional (Heath & Bizer, 2011).

2.4.8.2 Publicación de Linked Open Data en la Web

De acuerdo con los principios de Linked Data (T Berners-Lee, 2009), los proveedores de datos están obligados a añadir sus datos a un espacio de datos global para la publicación

de datos en la Web, lo que permite que los datos sean descubiertos y utilizados por varias aplicaciones. La publicación de un conjunto de datos como los datos relacionados en la Web incluye los siguientes pasos básicos (Christian Bizer, Heath, et al., 2009):

- Asignar URI's a las entidades descritas en el conjunto de datos y proporcionarles un modo de eliminación de referencias (URI's desreferenciadas) a través del protocolo HTTP y mediante representaciones basadas en RDF.
- Establecer enlaces basados en RDF hacia otras fuentes de datos en la Web con la finalidad de que los clientes naveguen por la Web de datos en su totalidad siguiendo los enlaces basados en RDF.
- Proporcionar metadatos acerca de los datos publicados, por lo que los clientes pueden evaluar la calidad de los datos publicados y elegir entre diferentes mecanismos de acceso.

La publicación de Linked Data requiere de la adopción de los cuatro principios de Linked Data (T Berners-Lee, 2009) (Véase sección 2.2.5). Sin embargo, su cumplimiento no implica el abandono de los sistemas de gestión de datos existentes y las aplicaciones de negocio, sino que simplemente es la adición de una capa técnica adicional para conectar éstos en la Web de datos (Heath & Bizer, 2011).

2.4.9 Linked Open Data cloud

Los orígenes de Linked Data, se encuentran en los esfuerzos de la comunidad de investigación de la Web Semántica y en particular en las actividades del proyecto del W3C *Linking Open Data cloud* (LOD cloud) (Heath, 2010). El objetivo inicial del proyecto fue generar una Web de datos mediante la identificación de conjuntos de datos existentes y disponibles bajo licencias abiertas *(Open Source Initiative)*[12], convertirlos a tripletas RDF de acuerdo con los principios de Linked Data (T Berners-Lee, 2009), y publicarlos en la Web. Como cuestión de inicial, el proyecto siempre ha estado abierto a cualquiera que publique datos de acuerdo con estos principios.

Los participantes involucrados en las etapas iniciales del proyecto fueron principalmente investigadores y desarrolladores en los laboratorios de investigación de diversas universidades y pequeñas empresas. Hoy en día, el proyecto ha crecido considerablemente incluyendo la participación significativa de grandes organizaciones como la BBC *(British Broadcasting Corporation)*, Thomson Reuters y la Biblioteca del

[12]Open Source Initiative: http://opensource.org/definition

Congreso de los Estados Unidos de América, por mencionar algunos. Este crecimiento es posible gracias al carácter abierto del proyecto, en el que cualquiera puede participar de maneta simple y mediante la publicación de un conjunto de datos de acuerdo con los principios de Linked Data (T Berners-Lee, 2009) y enlazándolos con los conjuntos de datos existentes (Christian Bizer, Heath, et al., 2009). La Figura 5, muestra el diagrama de la LOD cloud que representa un gran número de conjuntos de datos RDF interrelacionados que están siendo utilizados activamente por las comunidades de la industria, del gobierno y científicos (Curry, Freitas, & O'Riáin, 2010). Este diagrama representa conjuntos de datos publicados como Linked Open Data hasta Abril de año 2014.

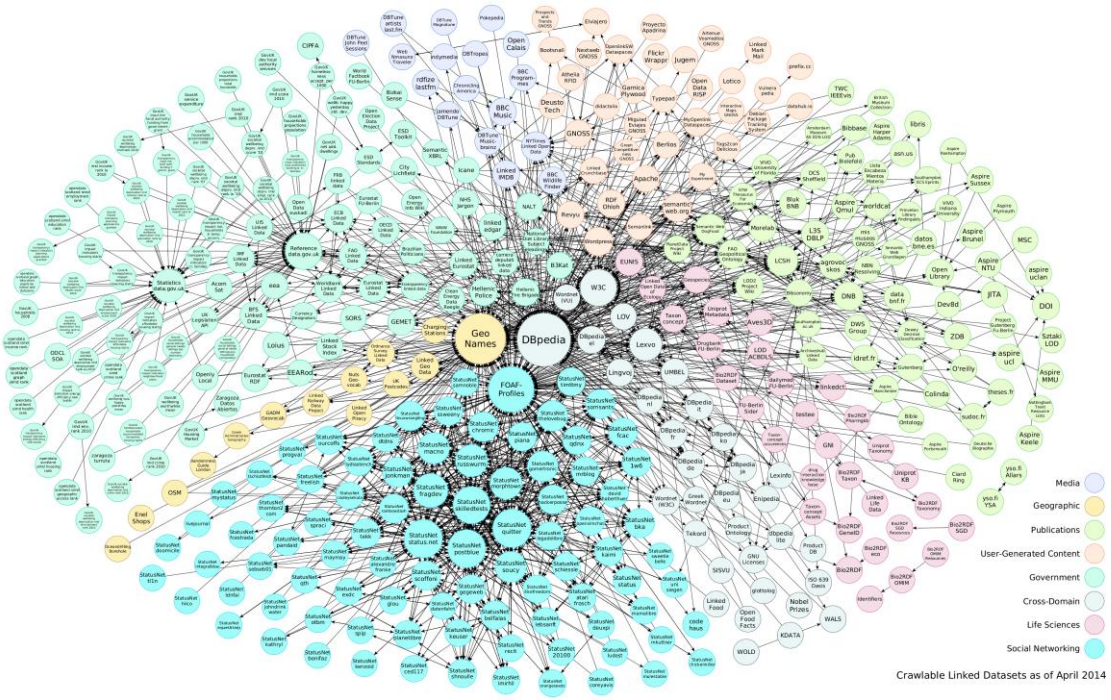


Figura 5. Diagrama de la Linked Open Data cloud (LOD cloud) hasta Abril de 2014 (Schmachtenberg, Paulheim, & Bizer, 2014)

La LOD cloud ha crecido considerablemente en los últimos años y proporciona una base sobre la cual las aplicaciones de Software aprovechan los datos disponibles en ella. A partir de Septiembre de 2011 consta de más de 300 conjuntos de datos abiertos con más de 35 mil millones de datos relacionados entre sí por 500 millones de enlaces con tipo (Bizer, Jentzsch, & Cyganiak, 2011). La LOD cloud está duplicando su tamaño cada 10 meses y está clasificada en diversos dominios entre los que se encuentran datos geográficos, datos de medios de comunicación e instituciones gubernamentales, así como de bibliotecas, educación, científicos y de salud, datos sobre ventas al pormenor y comercio, entre otros (Curry et al., 2013). Teniendo en cuenta el presente año (2014), y basándose únicamente en los dominios previamente mencionados, el número de conjuntos de datos en la LOD cloud

se ha duplicado desde 2011, sin embargo, añadiendo el dominio de redes sociales *(Social media)*, el número de conjuntos de datos se incrementa hasta un 271% (Schmachtenberg et al., 2014).

Cada conjunto de datos demuestra cómo la Web de datos enlazados está evolucionando, principalmente por la publicación de datos de terceros como usuarios entusiastas e investigadores, así como la publicación de datos a partir de fuentes de información proporcionadas por importantes medios de comunicación y organizaciones del sector público (Heath & Bizer, 2011). Se espera que esta tendencia tenga un impulso significativo, en organizaciones de otros sectores de la industria de modo que publiquen sus propios datos de acuerdo con los principios de Linked Data (T Berners-Lee, 2009).

2.4.9.1 DBpedia como núcleo de la Linked Open Data cloud

La Web tradicional está pasando de ser un medio para la conexión de documentos a un medio en el que también se comparten datos. Impulsado por la iniciativa *Open Data*, la Web de datos proporciona a los consumidores un fácil acceso a los datos disponibles en la LOD cloud (Curry et al., 2010). En este sentido, el núcleo de la LOD cloud es DBpedia (véase Figura 4), ya que ha derivado un corpus de datos a partir de la enciclopedia en línea Wikipedia. Esta última es muy visitada por los usuarios y mantiene una revisión constante. Las ediciones de Wikipedia están disponibles en más de 250 idiomas, tan sólo en Inglés hay 1,95 millones de artículos. Sin embargo, al igual que muchas otras aplicaciones Web, Wikipedia tiene problemas en sus capacidades de búsqueda que la limitan a realizar búsquedas de texto completo restringiendo el acceso a esta valiosa base de conocimientos (Auer et al., 2007). Wikipedia también tiene propiedades difíciles de tratar y que se presentan con los datos que son editados colaborativamente por ejemplo, su contenido es apenas interpretable por la máquina y carece de conocimiento estructural, por lo que se desconoce el cómo los conceptos se relacionan entre sí. Además, no permite ser declarado formalmente ni procesado automáticamente, la gran cantidad de datos numéricos está disponible sólo como texto y sin formato por lo tanto, no permite ser procesado por su significado real, tiene datos contradictorios, convenciones taxonómicas inconsistentes, errores, e incluso spam (Auer et al., 2007; Völkel et al., 2006).

El proyecto DBpedia se centra en la tarea de convertir el contenido de Wikipedia en conocimiento estructurado, de tal manera que se empleen contra él técnicas de la Web Semántica solicitando sofisticadas consultas sobre Wikipedia, vinculándolo con otros conjuntos de datos en la Web, creando nuevas aplicaciones de Software o utilizando y

combinando contenido de varias fuentes de datos *(mashups)* (Auer et al., 2007). La base de conocimientos DBpedia resultante describe hasta el año 2009, más de 2,6 millones entidades mientras que en el año 2007 la base de conocimientos constaba de 1,95 millones de entidades (Auer et al., 2007). Para cada una de estas entidades, DBpedia define un identificador único global cuya referencia a través de Internet permite ser anulada a través de una descripción RDF enriquecida de la entidad. Además, incluye definiciones legibles en 30 idiomas, relaciones con otros recursos, cuatro clasificaciones jerárquicas de conceptos así como enlaces a nivel de datos a otras fuentes de datos Web que describen a cada entidad (Bizer et al., 2009).

En los últimos años, un número creciente de proveedores de datos han comenzado a establecer vínculos a nivel de datos hacia los recursos DBpedia, haciendo de DBpedia un eje central de interconexión para la Web emergente de datos. Actualmente, la Web de fuentes de datos interrelacionados alrededor de DBpedia proporciona aproximadamente 4.7 billones de piezas de información y abarca ámbitos como información geográfica, personas, empresas, películas, música, genética, medicamentos, libros y publicaciones científicas (Bizer et al., 2009).

2.4.10 Repositorios RDF

Los repositorios de datos RDF son sistemas para almacenar y administrar datos descritos en RDF, también conocido como *RDF Store* o *Triple Store*. Algunos de los sistemas más utilizados se describen brevemente a continuación.

OpenLink Virtuoso Universal Server[13] es un sistema para la gestión de bases de datos de objeto-relacionales *(ORDBMS, Object Relational Database Management System)* y servidor de aplicaciones híbridas (también conocido como Universal Server) que proporciona la gestión, el acceso y la integración de datos. Cuenta con una versión empresarial la cual es de pago y la versión de código abierto disponible para la comunidad (Erling & Mikhailov, 2009). Las principales características de Virtuoso son:

- Gestión de Datos Relacionales
- Gestión de datos RDF
- Gestión de datos XML
- Gestión de indexación de contenidos en texto libre y de texto completo
- Servidor de documentos Web

[13]Virtuoso: http://virtuoso.openlinksw.com/

- Servidor de Linked Data
- Servidor de aplicaciones Web
- Despliegue de servicios Web (SOAP, Simple Object Access Protocol y REST, Representational State Transfer)

Dadas las características que OpenLink Virtuoso Universal Server ofrece y siguiendo los objetivos que con este trabajo de tesis se persiguen, resulta muy conveniente la implementación de la versión de código abierto ya que su servidor de Linked Data, proporciona capacidades suficientes para el almacenamiento y gestión de las tripletas RDF generadas a partir del modelo conceptual descrito en esta investigación.

El marco de trabajo Apache Jena (J. Carroll et al., 2004; Wilkinson et al., 2003), contiene un componente para el almacenamiento y consulta de datos RDF. Tal componente es nombrado TDB y admite la gama completa de las API de Apache Jena. TDB es utilizado como un almacén de datos RDF de alto rendimiento en una sola máquina y permite ser accedido y controlado mediante los scripts de línea de comandos proporcionados por la API de Apache Jena (Owens et al., 2008).

SESAME[14] es un marco de trabajo para el almacenamiento y gestión de datos RDF y vocabularios RDF(S) ampliamente utilizado en todo el mundo. Entre sus principales características destacan que permite distintos tipos de almacenamiento (ficheros o base de datos) y dar soporte a múltiples lenguajes de consulta, razonadores y protocolos cliente/servidor, en términos simpes, es una de las soluciones más flexibles disponibles para la gestión de datos RDF. En las versiones actuales de SESAME, se han desarrollado un gran número de extensiones para el RDF Store, tales como consultas adicionales SQL, consultas basadas en SPARQL, mejoras en la gestión, adaptadores de bases de datos y exportación, por mencionar algunos (Broekstra, Kampman, & Van Harmelen, 2002).

Bigdata RDF Database[15], es una base de datos de propósito general orientada a la escalabilidad horizontal, que permite ser desplegada a cientos de servidores. Entre sus características se encuentra la inclusión de una base de datos RDF de alto rendimiento que da soporte a RDF(S) y ofrece capacidades de inferencia basadas en reglas OWL Lite. Bigdata RDF Database soporta consultas basadas en SPARQL así como la indexación a través del uso eficiente la API *(Application Programming Interface)* de búsquedas Apache

[14]SESAME: http://www.openrdf.org/
[15]Bigdata RDF Database: http://www.systap.com/bigdata.htm

Lucene, entre muchas otras características. Cabe mencionar que su licencia es GNU *(General Public License)* lo que indica que es un producto para uso no comercial, en el caso de uso comercial es necesario adquirir una licencia de pago (B. Thompson & Personick, 2009).

En adición a los repositorios de datos RDF previamente descritos en esta sección, existen otras alternativas que tienen el propósito de almacenar, gestionar y proporcionar accesibilidad a los datos representados mediante tripletas RDF de la manera más óptima posible, algunas de estas alternativas son descritas brevemente a continuación:

- **4Store:** es una plataforma diseñada por Steve Harrys para respaldar sus aplicaciones basadas en Web Semántica. Tiene la capacidad de ejecutar consultas basadas en SPARQL sobre bases de datos de varios Terabytes y da soporte a una aplicación Web utilizada por miles de personas (Harris, Lamb, & Shadbolt, 2009).
- **OWLIM:** se trata de una familia de repositorios semánticos, o sistemas de gestión de bases de datos RDF con las siguientes características: motores nativos RDF, está implementado en Java, ofrece un rendimiento completo tanto en SESAME como en Apache Jena, ofrece un soporte robusto para la Semántica de RDF(S), OWL 2 RL y OWL 2 QL, posee una buena evaluación en su desempeño respecto a escalabilidad, carga y ejecución de consultas (Kiryakov, Ognyanov, & Manov, 2005).
- **RedStore:** es un RDF Triplestore ligero escrito en lenguaje C que utiliza el conjunto de bibliotecas RedLand. Cuenta con una interfaz HTTP y es compatible con los estándares SPARQL 1.1 del W3C (Humfrey, 2014).
- **HyperGraphDB:** es un mecanismo de código abierto y de propósito general para el almacenamiento de datos basado en un potente formalismo para la gestión del conocimiento conocido como híper-grafo dirigido. HyperGraphDB es una base de datos transaccional integrada e diseñada como un modelo de datos universal para aplicaciones altamente complejas que requieren una representación de conocimiento a gran escala como lo son Inteligencia Artificial, la bioinformática y el procesamiento del lenguaje natural (Iordanov, 2010).
- **AllegroGraph:** es un gestor moderno con alto rendimiento en la persistencia de bases de datos basadas en grafos, hace uso eficiente de la memoria en combinación con el almacenamiento basado en disco, habilitándolo con capacidades de escalabilidad para miles de millones de tripletas mientras mantiene un alto rendimiento en su funcionalidad. Finalmente, AllegroGraph da soporte a SPARQL,

RDFS++ y al razonamiento basado en Prolog a partir de numerosas aplicaciones cliente (Aasman, 2006; Rohloff et al., 2007).

El uso de los repositorios RDF es crucial para la gestión y el almacenamiento de los datos descritos en alguna notación semántica, y que se quiere, sean puestos a disposición de los usuarios para ser explotados de la marera que resulte más conveniente. Sin embargo, para que estos repositorios almacenen a este tipo de datos, es necesario que estos sean procesados y transformados, para ello existe una diversidad de herramientas que permiten la transformación y generación de grafos RDF a partir de fuentes de información estructurada, semiestructurada e incluso no estructurada. Algunas de esas herramientas son descritas en la sección siguiente.

2.4.11 Herramientas para la publicación de Linked Data

Actualmente se han desarrollado una variedad de herramientas para la publicación de Linked Data. Estas herramientas sirven, ya sea, para mostrar el contenido almacenado en los repositorios RDF como Linked Data en la Web, o para proporcionar vistas Linked Data a partir de fuentes de datos existentes que no son RDF. Además, este tipo de herramientas también nombradas *Linked Data Frontend*, apoyan a los usuarios con el tratamiento de los detalles técnicos tales como, la transacción de contenidos y a que se aseguren de que los datos son publicados de acuerdo a las mejores prácticas establecidas por la comunidad Linked Data (Berrueta & Phipps, 2008; Sauermann, Cyganiak, & Völkel, 2008). Todas las herramientas para la publicación de Linked Data, dan soporte a las URIs desreferenciadas descritas en los RDF. Finalmente, algunas de estas herramientas también proporcionan acceso a los conjuntos de datos publicados a través de consultas basadas en SPARQL y apoyan la publicación de los vertederos de RDF (Bizer, Cyganiak, & Heath, 2007). Algunas herramientas de esta categoría se describen brevemente a continuación.

El previamente descrito OpenLink Virtuoso Universal Server (Erling & Mikhailov, 2009), ofrece una vista de los datos RDF almacenados en él a través de una interfaz Linked Data y un SPARQL endpoint sin embargo, puede ser complementado a través de Pubby Server. Pubby permite ser utilizado como una extensión de cualquier repositorio RDF que dé soporte a SPARQL. Entre sus características, se encuentra el uso de expresiones regulares para definir patrones de URI ya que Pubby, reescribe solicitudes de URI en consultas basadas en SPARQL-DESCRIBE para el repositorio RDF subyacente. Además de RDF, Pubby proporciona una vista simple en HTML sobre el repositorio de datos y se

encarga de manejar las redirecciones HTTP 303 y la negociación de contenidos entre las dos representaciones (Cyganiak & Bizer, 2008).

D2R Server[16] forma parte de la Plataforma D2RQ, fue desarrollado por el Grupo de Sistemas basados en la Web de la Freie Universität de Berlín. DR2 Server permite publicar el contenido de las bases de datos relacionales en la Web Semántica siguiendo el formato RDF, proporciona un visualizador RDF y HTML para examinar el contenido de las bases de datos además, ofrece a las aplicaciones la capacidad de consultas las bases de datos mediante consultas basadas en SPARQL (Christian Bizer & Cyganiak, 2006). Por otra parte, la Plataforma Talis, se ofrece como un producto de Software como servicio *(Software as a service)* al que se accede a través del protocolo HTTP y proporciona almacenamiento nativo para RDF/Linked Data. Sus derechos de acceso permiten que los contenidos de cada repositorio RDF sean accesibles a través de un SPARQL endpoint y una serie de APIs REST que se ajustan a los principios de Linked Data (T Berners-Lee, 2009; Bizer, et al., 2009; P. Miller, Styles, & Heath, 2008).

WESO DESH es un potente Linked Data Frontend empresarial para SPARQL endpoints desarrollado en Java que proporciona una forma configurable de acceder a datos RDF utilizando URL's RESTful simples que se traducen en consultas para un SPARQL endpoint. Sus principales características son la negociación de contenido mediante el protocolo HTTP 303, manejo de caché de resultados, salida HTML + RDFa nativo, URIs definidas a través de expresiones regulares, la ejecución de múltiples tipos de consultas basadas en SPARQL (CONSTRUCT, ASK and DESCRIBE), administración a través de una interfaz grafica de usuario y permitir patrones complejas de URIs utilizando expresiones regulares (Alvarez et al., 2012).

El kit de herramientas Triplify, apoya a los desarrolladores en la extensión de aplicaciones Web existentes mediante Linked Data Frontends. Esta extensión se efectúa a partir de la implementación de un conjunto de plantillas basadas en consultas SPARQL, por lo que este kit de herramientas proporciona una vista en Linked Data y en JSON sobre la base de datos de las aplicaciones extendidas (Auer et al., 2009). Siguiendo en el contexto de los servicios dirigidos hacia las aplicaciones Web, SparqPlug es un servicio que permite la extracción de datos Linked Data a partir de documentos HTML que no contienen datos RDF. El servicio de SparqPlug funciona por la serialización del DOM *(Document Object Model)* del HTML como RDF y permitiendo a los usuarios definir consultas basadas en

[16]D2R Server: http://d2rq.org/d2r-server

SPARQL que transforman elementos de este en un grafo de su elección (Coetzee, Heath, & Motta, 2008).

Entre las herramientas para la publicación de Linked Data también se incluyen envoltorios o *wrappers*, dos ejemplos de ellos son OAI2LOD Server y SIOC Exporters. El primero es un envoltorio Linked Data para servidores de documentos que dan soporte al protocolo OAI-RMH (Haslhofer & Schandl, 2008) y el segundo consiste en varios envoltorios de Linked Data para varios motores de blogs populares y sistemas de gestión de contenidos y foros de discusión tales como WordPress, Drupal y phpBB (Bojars et al., 2008).

2.4.12 Herramientas para el descubrimiento de enlaces RDF en la Web de datos

En los últimos años, la publicación de conjuntos de datos que siguen los principios de Linked Data (T Berners-Lee, 2009) procedentes de una amplia gama de dominios, se está volviendo cada vez más importante para la apertura de la información (Sánchez-Cervantes et al., 2013). Complementario a esto, cada vez se desarrollan más herramientas disponibles para la publicación de Linked Data, al mismo tiempo que surgen herramientas para el descubrimiento de enlaces RDF entre diferentes fuentes de datos contenidas en la LOD cloud. En este sentido, algunas de estas herramientas se describen brevemente a continuación:

- **Silk Link Discovery Framework:** es un conjunto de herramientas para el descubrimiento y el mantenimiento de enlaces de datos contenidos entre las distintas fuentes existentes en la Web (Volz et al., 2009). Silk apoya a los publicadores de datos en el establecimiento de enlaces RDF explícitos, de manera que estos tienen la posibilidad de especificar cuáles enlaces RDF deben ser descubiertos entre las fuentes de datos y qué condiciones deben cumplir para ser interconectados. Estas condiciones permiten combinar varias métricas de similitud y permiten tomar el grafo RDF entorno a un elemento de datos especificado, el cual es dirigido utilizando una ruta en lenguaje RDF. Silk accede a las fuentes de datos idóneas a ser vinculadas entre sí a través del protocolo SPARQL y por lo tanto, permiten ser utilizadas de manera local o de manera remota mediante un SPARQL endpoint. Silk es utilizable a través de la interfaz gráfica de usuario Silk Workbench o desde línea de comandos. Ambas variantes se basan en el motor de descubrimiento de enlaces Silk que ofrece las siguientes características (Volz et al., 2009a):

  - Ofrece un lenguaje flexible y declarativo para especificar reglas de vinculación.
  - Proporciona soporte para la generación de enlaces RDF (como *owl:sameAs* y enlaces de otros tipos).
  - Facilita su empleo en entornos distribuidos (mediante el acceso a los SPARQL enpoints locales y remotos).
  - Da soporte para su uso en situaciones en los que términos de diferentes vocabularios se mezclan y donde no existen esquemas RDF(S) u OWL consistentes.

- **LIMES:** es un marco de trabajo para el descubrimiento de enlaces en la Web de Datos. LIMES *(LInk discovery framework for MEtric Spaces)* implementa métodos eficientes en tiempo para el descubrimiento a gran escala de enlaces basados en las características de los espacios métricos. Además, es de fácil configuración a través de una interfaz Web y se puede descargar como herramienta independiente para llevar a cabo el descubrimiento de enlace a nivel local. LIMES consta de siete módulos principales de los cuales cada uno puede ser ampliado para dar cabida a funcionalidades nuevas o mejoradas. Los módulos centrales de LIMES son un Módulo controlador, que coordina el proceso de correspondencia y un módulo de datos, que contiene todas las clases necesarias para almacenar datos. El proceso de comparación se lleva a cabo de la siguiente manera: en primer lugar, el módulo controlador invoca al módulo de entradas y salidas (I/O module), que lee el fichero de configuración y extrae toda la información necesaria para llevar a cabo la comparación de los casos, incluyendo las URL's de los SPARQL endpoints de la bases de conocimiento, la expresión de la métrica que se utilizará y el umbral en el que se utilizarán las restricciones a las instancias de mapeo (Por ejemplo, su tipo) (Ngomo & Auer, 2011).

La Web de datos se basa en dos ideas sencillas: en primer lugar, se emplea el modelo de datos RDF para la publicación de datos estructurados en la Web y en segundo lugar, se establecen los enlaces RDF explícitos entre los elementos de datos dentro de las diferentes fuentes de datos (Volz et al., 2009a). Con base en estos argumentos, las herramientas como Silk y LIMES son indispensables para el descubrimiento e interconexión de datos contenidos en fuentes externas, con los almacenados en la base de conocimientos financieros basada en Linked Data que se describe en el Capítulo 4.

2.4.13 Lenguajes para la consulta de grafos RDF

El RDF es considerado el estándar de mayor relevancia para la representación e intercambio de datos en la Web Semántica, la estructura que subyace a cualquier documento RDF es una colección de tripletas también conocida como grafo RDF. Estos grafos se sustentan por el modelo abstracto de datos RDF, el cual es independiente de una sintaxis de serialización concreta, esto significa que por lo general, los lenguajes de consulta para este tipo de grafos no presentan características para consultar funciones de serialización específicas (Haase et al., 2004). En los párrafos siguientes se proporciona una breve descripción acerca de los principales lenguajes para la consulta de grafos RDF.

TRIPLE, el término *"Triple"* denota tanto una consulta como un lenguaje de reglas, así como el tiempo de ejecución real del sistema. Este lenguaje se deriva de F-Logic (Balaban, 1995) y las tripletas RDF son representadas como expresiones F-Logic anidadas. TRIPLE no codifica una semántica RDF fija. La semántica deseada tiene que especificarse como un conjunto de reglas, junto con la consulta. Además, los tipos de datos no son compatibles con TRIPLE (Sintek & Decker, 2002).

RQL *(Resource Query Language)* en un lenguaje para la consulta de datos RDF que sigue un enfoque funcional que soporta expresiones de ruta generalizadas ofreciendo variables en ambos nodos y artistas del grafo RDF consultado. RQL se basa en un modelo formal que captura las primitivas del modelo del grafo RDF y permite la interpretación de descripciones de recursos superpuestos por medio de uno o más esquemas. La novedad de RQL radica en su capacidad de combinar sin problemas el esquema y los datos de la consulta, mientras que explota la taxonomía de etiquetas y la clasificación múltiple de los recursos. RQL sigue una sintaxis OQL-Like, es ortogonal pero no limitado como las consultas que devuelven enlaces de variables en lugar de grafos. Sin embargo, la semántica de RQL no es totalmente compatible con la Semántica RDF, una serie de restricciones adicionales son colocadas en los modelos RDF para permitir las consultas con RQL (Karvounarakis et al., 2002).

SquishQL es un lenguaje simple para la consulta de datos RDF, en él, un termino Squish es tratado como "SQL-ish" y la estructura básica de su sintaxis está diseñada para parecerse a la estructura de SQL en la que a través de una consulta realizada a una base de datos, se obtienen los valores para una selección de variables dada en una expresión limitante. En el contexto de RDF, los términos SQL-ish proporcionan el acceso a repositorios semánticos y grandes bases de datos de maneta consistente y comprensible

para el humano permitiendo a los desarrolladores de aplicaciones de la Web Semántica rápidamente (Miller, Seaborne, & Reggiori, 2002).

A partir de SquishQL se derivó RDQL *(RDF Data Query Language)*, que fue desarrollado por Hewlett Packard, presentado en el W3C en Enero de 2004 y forma parte del conjunto de herramientas Apache Jena RDF así como de una serie de sistemas tales como RDFStore, SESAME, clases PHP XML, 3 Store y las APIs RAP-RDF de PHP para la extracción de información a partir de grafos RDF (Seaborne, 2004). Por otra parte, la sintaxis de RDQL sigue un patrón de selección *(Select)* similar a SQL, en el que la cláusula *from* es omitida (Haase et al., 2004).

SerQL está implementado y disponible en el sistema SESAME, se trata de un lenguaje para la consulta de datos RDF, está basado en los principios de varios lenguajes como RQL, RDQL y N3 y como tal, representa un lenguaje de segunda generación (Broekstra & Kampman, 2003). Sus principales objetivos de diseño son la unificación de las mejores prácticas del lenguaje de consultas y la entrega básica de un lenguaje de con toda la expresividad para consultas de datos RDF que se ocupa de las situaciones prácticas. La sintaxis de SeRQL es similar a la de RQL aunque se han hecho modificaciones para hacer que el lenguaje sea más fácil de analizar. Como RQL, SeRQL se basa en una interpretación formal del grafo RDF, pero la interpretación oficial de SeRQL se basa directamente en la teoría de modelos de RDF (Broekstra & Kampman, 2004). Cabe destacar que otra característica interesante de SeRQL es que representa una colaboración entre la industria y la comunidad de código abierto y aparentemente ofrece una alternativa viable a la estandarización de consultas RDF basándose en el proceso del W3C (Hutt, 2005).

Un lenguaje para la consulta de datos RDF con un enfoque interesante es Versa, su principal elemento es una lista de recursos RDF. En este sentido, las tripletas RDF juegan un importante papel en las invocaciones a las operaciones transversales que recorren a las tripletas RDF siguiendo la forma siguiente ListExpr – ListExpr -> boolexpr. Estas expresiones devuelven una lista de todos los objetos de las tripletas coincidentes. Versa tiende a apoyar las reglas, ya que permite atravesar los predicados de manera transitiva, pero no permite la implementación de vistas, múltiples modelos y manipulación de datos. Sin embargo, Versa cumple los criterios de ortogonalidad y seguridad (Haase et al., 2004; Ogbuji, 2005). Además, este lenguaje es apoyado por 4Suite, que es un conjunto de herramientas XML y RDF (Chagas, De-Carvalho, & Da-Silva, 2008).

XsRQL es un lenguaje de consultas de datos RDF que obtiene la mayor parte de su sintaxis y estilo a partir de X-Query al tiempo de aprovechar muchas de las características útiles desarrolladas por el Grupo de Trabajo XML Query del W3C omitiendo partes de la especificación X-Query que son específicas de XML y que por lo tanto no se requiere en un entorno RDF (Katz, 2004). En adición, XsRQL es un lenguaje que aprovecha los enfoques actuales de XML, lo que da una idea de cómo se pueden agregar consultas de datos RDF como una extensión a otras tecnologías (Hutt, 2005).

El acceso integrado a múltiples fuentes de datos RDF distribuidas y autónomas es un desafío clave para muchas aplicaciones de la Web Semántica (Hutt, 2005). Como reacción a este desafío, se desarrolló el lenguaje SPARQL *(Simple Protocol and RDF Query Language)*, el cual cumple los requisitos y objetivos de diseño descritos en el la recomendación *"RDF Data Access Use Cases and Requirements"* del *W3C RDF Data Access Working Group* (DAWG) para proporcionar apoyo a la consulta de múltiples grafos RDF (Clark, 2005). Esto significa que SPARQL proporciona la capacidad para expresar consultas a través de diversas fuentes de datos, siempre y cuando los datos se almacenen en forma de RDF o sean definidos como vistas a través de un sistema intermediario *(Middleware)*. Las capacidades de SPARQL permiten realizar consultas en los patrones obligatorios y opcionales del grafo RDF consultado, incluyendo sus conjunciones y disyunciones incluyendo el soporte para la ampliación o aplicación de restricciones en el ámbito de las consultas indicando los grafos sobre los que se opera. Los resultados de las consultas SPARQL son conjuntos de resultados o grafos RDF (Prud'Hommeaux & Seaborne, 2008). Una consulta SPARQL consta de tres partes: a) La parte de coincidencia de patrones *(Pattern matching)*, que incluye varias características interesantes en la coincidencia de patrones de los grafos tales como partes opcionales, unión de patrones, anidación, filtrado (o restricciones) y valores susceptibles a ser emparejados y la posibilidad de elegir la fuente de datos a ser emparejada por un patrón; b) Los modificadores de solución *(Solution modifiers)*, los cuales una vez que la salida del modelo se ha calculado (En forma de tabla con los valores de las variables), permiten modificar estos valores modificando esos valores aplicando los operadores clásicos como like, projection, distinct, order, limit, y offset; y c) Los resultados de la consulta SPARQL *(SPARQL results)*, estos se presentan de diferentes maneras: como consultas booleanas (Yes/No queries), la selección de los valores de las variables que se ajustan a los patrones, como la construcción de nuevas tripletas a partir de los valores obtenidos y como la descripción de los recursos (Pérez, Arenas, & Gutierrez, 2006). SPARQL ha evolucionado con el tiempo por ejemplo, hasta el momento el estándar de

SPARQL 1.1 presenta un conjunto de especificaciones que proporcionan lenguajes y protocolos para consultar y manipular grafos RDF contenidos en la Web o en algún RDF Store. Este estándar comprende las siguientes especificaciones (W3C, 2013):

- Lenguaje de consulta para RDF *(RDF Query language)*.
- Despliegue de resultados en los formatos JSON, CSV *(Comma Separated Values)* y TSV *(Tab Separated Values),* además de continuar con el formato XML del estándar SPARQL 1.0.
- Consultas federadas *(Federated Query)*, se trata de una especificación que define una extensión del lenguaje SPARQL 1.1 para ejecutar consultas distribuidas en diferentes SPARQL endpoints.
- Regímenes de vinculación *(Entailment Regimes)*, es una especificación que define la semántica de las consultas SPARQL bajo regímenes de vinculación como RDF(S), OWL o RIF *(Rule Interchange Format)* (Kifer & Boley, 2013).
- Lenguaje de actualización para grafos RDF.
- Protocolo para RDF, se trata de un protocolo que define los medios para transmitir consultas SPARQL arbitrarias y actualizar solicitudes a un servicio de SPARQL.
- Descripción de servicios, es una especificación que define un método para descubrir un vocabulario y para describir servicios SPARQL.
- Protocolo HTTP para el almacenamiento de grafos *(Graph Store HTTP Protocol)*, a diferencia del protocolo full SPARQL, esta especificación define los medios para la gestión del contenido de los grafos RDF directamente a través de operaciones HTTP comunes.
- Casos de prueba *(Test Cases)*, consiste en una suite de pruebas útil para la comprensión, descripción y evaluación de casos en los que se requiere saber si un sistema está desarrollado conforme al estándar SPARQL 1.1

El uso de tecnologías semánticas favorece en gran medida la publicación de datos basados en el paradigma Linked Data, sin embargo, el proceso de publicación bajo dicho paradigma requiere del seguimiento de un ciclo de vida que contenga determinadas fases y que sirva de metodología común para delimitar, organizar, documentar e identificar las funciones que se realizan con los datos que están siendo publicados. Dicho esto, en la sección siguiente se proporciona información referente a los Ciclos de vida de Linked Data.

2.4.14 Ciclos de vida de Linked Data

El paradigma de Linked Data ha evolucionado como un poderoso habilitador para la transición de la Web tradicional orientada a documentos, en una Web de datos relacionados entre sí y, en última instancia, en la Web Semántica. El término Linked Data se refiere a un conjunto de mejores prácticas para la publicación y conexión de datos estructurados en la Web. El proceso de publicación de Linked Data, sigue determinados ciclos de vida de la misma manera que la Ingeniería de Software. Algunos autores como (Hyland et al., 2011) y (M Hausenblas et al., 2011) los representan como una secuencia de etapas no iterativa, tal y como se muestran en las Figuras 6 y 7.

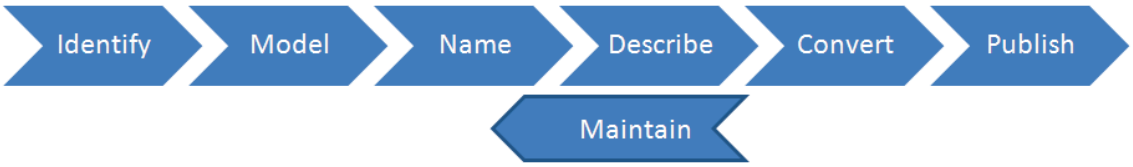


Figura 6. Ciclo de vida de Linked Data de acuerdo a Hyland et al., (2011)

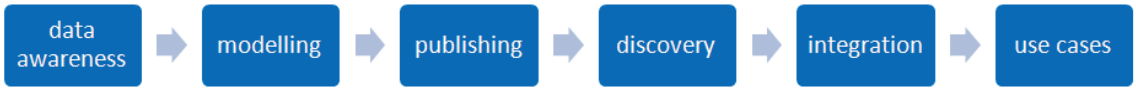


Figura 7. Ciclo de vida de Linked Data de acuerdo a M Hausenblas et al., (2011)

Por otra parte, el proceso de publicación del *Goverment Linked Data* debe tener un ciclo de vida, de la misma manera de la Ingeniería del Software, en la que cada proyecto de desarrollo tiene un ciclo de vida, de acuerdo con la experiencia de (Villazón-Terrazas et al., 2011), este proceso tiene un modelo de ciclo de vida iterativo gradual, que se basa en la mejora y ampliación del *Goverment Linked Data* que resulta de la realización de varias iteraciones. La Figura 8, muestra un ejemplo de este tipo de proceso.

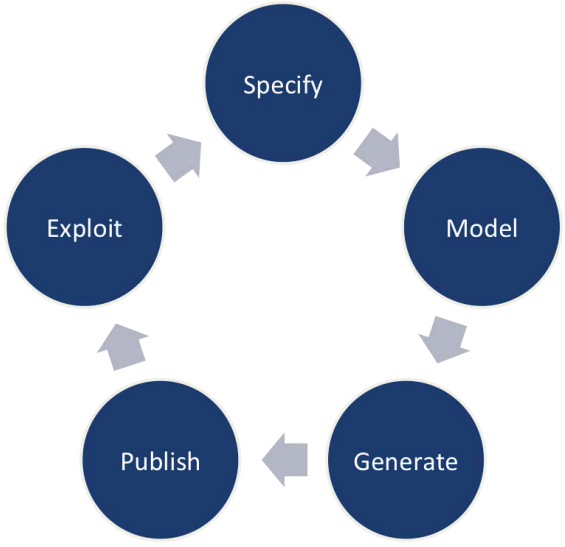


Figura 8. Ciclo de vida de Linked Data iterativo de acuerdo a Villazón-Terrazas et al., (2011)

Este ciclo de vida, implica las siguientes tareas y actividades:

- Especificación
- Diseño
- Definición/Descripción de la procedencia de la información
- Análisis de las fuentes de datos
- Modelado
- Búsqueda de ontologías adecuadas/vocabularios que modelan las fuentes de datos
- Creación del modelo mediante la reutilización de las ontologías/ vocabularios seleccionados
- Generación
- Transformación del origen de datos RDF
- Limpieza de datos
- Enlaces (Vínculos)
- Identificación de los conjuntos de datos que pueden ser adecuados como objetivos de vinculación
- Descubrimiento de las relaciones entre los elementos de datos del *Government Dataset* y los elementos de los conjuntos de datos identificados en el paso anterior
- Validación las relaciones que se han descubierto en el paso anterior
- Publicación
- Publicación del conjunto de datos
- Publicación de metadatos
- Habilitación de la detección efectiva
- Explotación

Otro ciclo de vida iterativo para la publicación de Linked Data (véase Figura 9) es el presentado por (Auer et al., 2013), para el proyecto LOD2[17], cuya descripción general se proporciona a continuación:

[17] LOD2: http://lod2.eu/Welcome.html

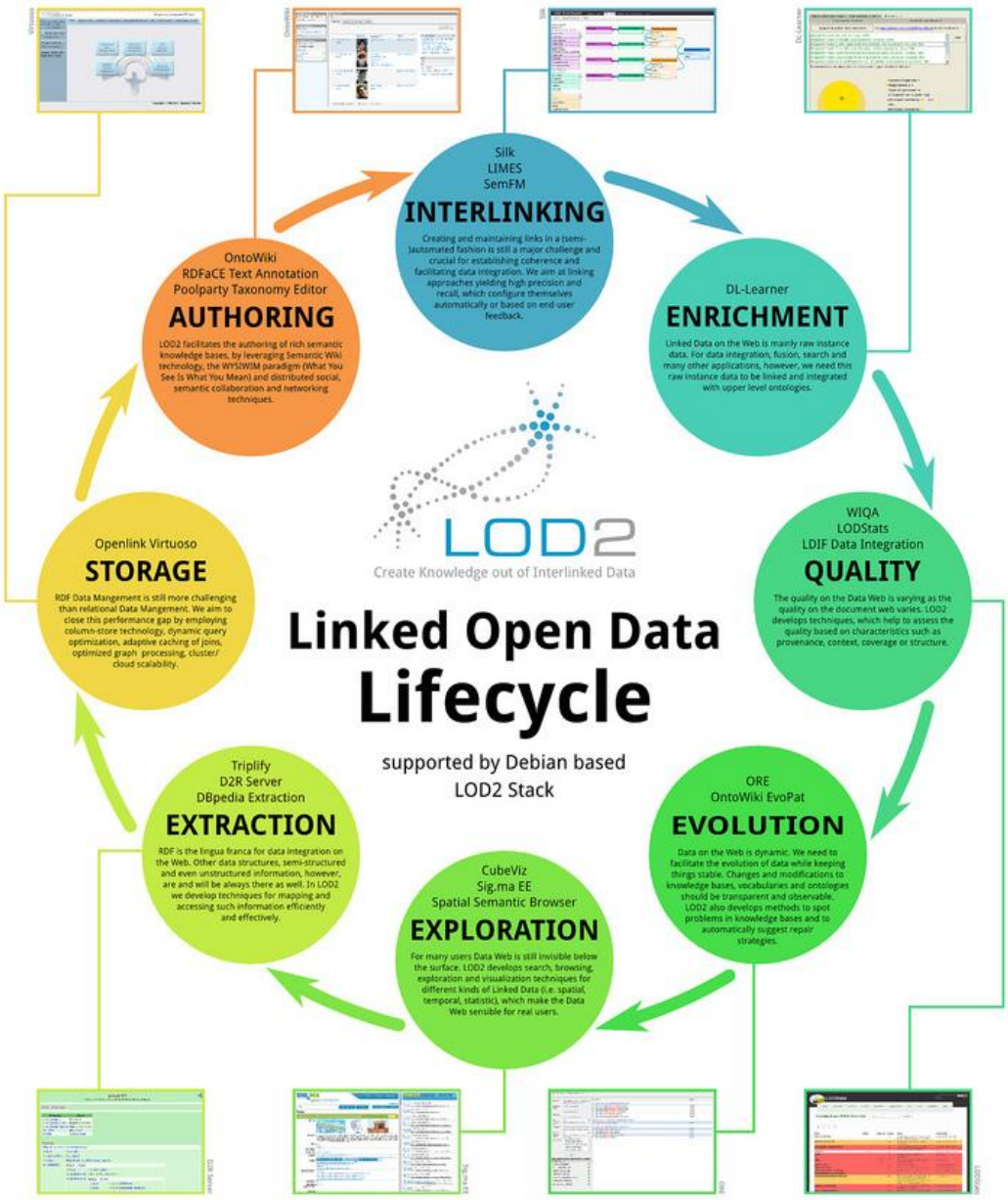


Figura 9. Ciclo de vida de Linked Data para el proyecto LOD 2 (Auer et al., 2013)

1. **Extracción *(Extraction)*:** la información representada en forma no estructurada, o que se adhiera a otros formalismos de representación de datos estructurados o semiestructurados, debe ser comparada con el modelo de datos RDF. El uso de herramientas como Triplify, D2R Server y DBpedia Extraction son útiles en esta fase.
2. **Almacenamiento y consultas *(Storage & Querying)*:** una vez que hay una masa crítica de datos RDF, los mecanismos tienen que estar en su lugar para almacenar, indexar y consultar estos datos RDF eficientemente. Una de las herramientas que facilitan el almacenamiento y consultas es Openlink Virtuoso.

3. **Autoría *(Authoring)*:** los usuarios deben tener la oportunidad de crear nueva información estructurada o de corregir y ampliar la ya existente. Las herramientas posibles a utilizar en esta etapa son OntoWiki, RDFaCE Text Annotation y Poolparty Taxonomy Editor.
4. **Interconexión de datos *(Interlinking)*:** si los diferentes proveedores de datos proporcionan información sobre los mismos o entidades relacionadas, los enlaces entre los diferentes activos de información deben ser establecidos. Herramientas como Silk, LIMES y SemFM, facilitan la interconexión de datos en esta fase.
5. **Enriquecimiento *(Enrichment)*:** desde que Linked Data comprende principalmente datos de instancia, se observa una falta de clasificación, estructura y esquema de información. Esta deficiencia puede ser abordada por enfoques para el enriquecimiento de datos con estructuras de nivel superior con el fin de ser capaz de agregar y consultar los datos de manera más eficiente. Una herramienta útil en esta fase es DL-Learner.
6. **Análisis de la calidad *(Quality Analysis)*:** al igual que con los documentos de la Web tradicional, la Web de datos contiene una variedad de información de diferente calidad. Por lo tanto, es importante el diseño de estrategias para evaluar la calidad de los datos publicados en esta. Algunas herramientas que sirven de apoyo en esta fase son WIQA, LODStats, LDIF Data Integration.
7. **Evolución y reparaciones *(Evolution & Repair)*:** una vez que se detectan problemas, son requeridas estrategias para la reparación de estos problemas y apoyar la evolución de Linked Data. Las herramientas ORE y OntoWiki EvoPat, sirven de apoyo en esta fase.
8. **Búsqueda, navegación y exploración *(Search, Browsing & Exploration)*:** por último, pero no menos importante, los usuarios tienen que estar capacitados para navegar, buscar y explorar la información de la estructura disponible en la Web de datos de una manera rápida y amigable con el usuario. CubeViz, Sigma EE y Spatial Semantic Browser, son herramientas que sirven de apoyo en esta fase.

Las diferentes etapas del ciclo de vida de Linked Data no existen en forma aislada o se pasan (iteran) en una secuencia estricta, pero mutuamente, se enriquecen a sí mismas. (Auer et al., 2013).

Visto como un ascensor de datos *(DataLift Vision)*, el proceso de publicación de Linked Data se divide en dos fases principales, la fase de Apertura de los datos y la fase de

Publicación del conjunto de datos (M Hausenblas et al., 2011). La Figura 10, muestra este proceso de publicación de Linked Data.

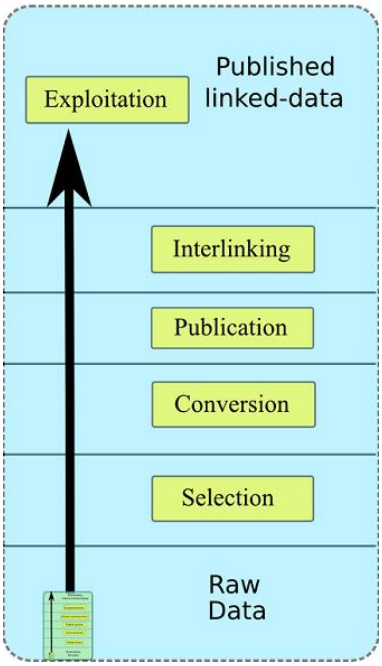


Figura 10. Proceso de publicación de Linked Data visto como un ascensor de datos

La fase de apertura de los datos consiste en (M Hausenblas et al., 2011):

- Apoyar en la selección de datos
- Identificar el vocabulario relevante
- Definir un modelo de esquema para los URIs
- Convertir entre formatos
- Almacenar los datos RDF en un almacén de tripletas *(Triple Store)*
- Interconectar los datos

La fase de publicación del conjunto de datos se compone de:

- Adjuntar la procedencia de los datos
- Gestionar los derechos de acceso al conjunto de datos

Proporcionar información concerniente a los fundamentos financieros y tecnológicos que subyacen a esta tesis proporciona más a detalle una comprensión acerca del modelo semántico inspirado en Linked Data que se presenta en la misma. Sin embargo, es necesario el análisis de otras investigaciones realizadas por diversos autores a fin de

identificar aciertos y deficiencias que sirvan de punto de comparación con esta investigación. Por este motivo, en la sección siguiente se continua con la descripción de varios trabajos relacionados con esta tesis.

2.5 Trabajos relacionados

Hay varias iniciativas relacionadas con esta investigación, los trabajos relacionados que se describen a continuación están clasificados en dos partes. La primera parte incluye obras basadas en la extracción, procesamiento y representación de datos a partir de fuentes de información financiera. En la segunda parte, se describen las iniciativas relacionadas con el análisis de la información financiera. Algunas de estas obras han obtenido buenos resultados, mientras otras sólo han obtenido resultados parciales. En comparación con la investigación llevada a cabo en esta tesis, estos trabajos presentan diferentes objetivos, métodos y técnicas utilizadas para llegar a sus conclusiones.

2.5.1 Extracción, procesamiento y representación de datos financieros

La presentación estructurada, pero la débil interconexión de los datos contenidos en documentos XBRL, los hace candidatos interesantes para ser transformados en forma de Linked Data (García & Gil, 2010a). De acuerdo con lo expresado, La tecnología Linked Data fue creada para la integración de información a escala Web, provee el alojamiento de datos XBRL y facilita su combinación con otros conjuntos de datos abiertos disponibles en la Web (O'Riain, Curry, et al., 2012). Un ejemplo de esto, es la transformación de 1000 documentos XBRL en forma de Linked Data que dio como resultado la obtención de más 3 millones de tripletas RDF (García & Gil, 2010b). La importancia de este trabajo para esta investigación, consiste en que proporciona un punto de partida para la adquisición y transformación de documentos XBRL en forma de Linked Data ya que, en este trabajo de tesis, se generaron 138,664,094 millones de tripletas RDF a partir de 830,321 documentos XBRL extraídos de la EDGAR, correspondientes al periodo del segundo trimestre del año 2008 al cuarto trimestre del año 2013 (y continua incrementándose).

Las taxonomías financieras y los marcos de trabajo basados en ontologías mejoran la calidad de los datos financieros en línea (Du & Zhou, 2012). La confirmación de esta aseveración fue el trabajo de estos autores. En él, proporcionan una alternativa de solución a los problemas de ambigüedad, inconsistencia, inexactitud, falsedad y de valores incompletos presentes en los datos financieros en línea. En el trabajo de esta tesis se integran los datos financieros bajo las taxonomías financieras correspondientes a las Hojas

de Balance, Cuenta de resultados y Estado de flujos de efectivo, basadas en la norma US-GAAP con el fin de proporcionar una base de conocimientos financieros que siga los principios de Linked Data (T Berners-Lee, 2009) y que sirva como una alternativa de solución a los problemas descritos anteriormente, además de ayudar a mejorar la calidad de los datos financieros.

Los Estados financieros son vitales para los tomadores de decisiones en el mundo de la inversión profesional, donde el acceso rápido y la importación automática de datos es esencial para la industria financiera. La fuente más común de información financiera es el repositorio de datos EDGAR a través de los documentos XBRL almacenados en ella. En términos técnicos, los datos disponibles en estos documentos están débilmente estructurados por el hecho de almacenar contenido Mixto, es decir, tanto el texto en lenguaje natural como los datos financieros, se presentan de manera semiestructurada en forma de tablas (Stümpert, 2008). El modelo semántico que se presenta en esta tesis, mejora la presentación de la información financiera contenida en el repositorio de datos EDGAR a través de la extracción y el procesamiento de la información almacenada en los documentos XBRL que ahí se albergan. El resultado es una base semántica de conocimientos financieros inspirada en los principios de Linked Data (T Berners-Lee, 2009).

Los aspectos técnicos de los documentos XBRL y los esfuerzos de la SEC para incorporar este tipo de documentos en la presentación de datos financieros, es un enfoque *centrado en los datos* que resulta adecuado para reemplazar el actual paradigma orientado a la presentación de Estados financieros en documentos de papel. Sin embargo, este enfoque carece de la interconexión e interoperabilidad entre los datos (Plumlee & Plumlee, 2008). Este inconveniente se resuelve con la iniciativa presentada en este trabajo de tesis, a través de las taxonomías financieras (Hojas de Balance, Cuenta de resultados y Estado de flujos de efectivo) que son reutilizables por otros sistemas. Además, el hecho de que todos los conceptos almacenados en la base de conocimientos financieros se representen utilizando URI's relacionables entre sí y que sean fácilmente recuperados por otros sistemas mediante el procesamiento de consultas basadas en SPARQL, es una muestra de una notable mejora en la interoperabilidad de los datos.

Núñez et al., (2008) presentan un proceso para la construcción de modelos de información explícitos para los fondos de inversión en el mercado Español. El proceso presentado por los autores, incluye la traducción de las taxonomías XBRL en OWL *(Web*

Ontology Language) que permite el intercambio y análisis de la información de los fondos de inversión. Los autores argumentan que la falta de modelos explícitos y compartidos para el intercambio de información en el mercado de fondos de inversión así como la promoción y la creciente adopción del estándar XBRL en España por parte de los reguladores y supervisores como el Banco de España[18] y la CNMV[19] *(Comisión Nacional del Mercado de Valores)* les llevó a considerar a XBRL como lenguaje candidato para la creación de un modelo de información financiera explícita. Núñez y sus colaboradores identificaron a OWL como una alternativa potencial para el uso de XBRL ya que presenta algunas características de interés práctico para ser aprovechadas en el mercado de fondos de inversión. Con el fin de evaluar el uso de ontologías OWL, ellos desarrollaron un proceso de traducción genérica de taxonomías XBRL en ontologías OWL para que las taxonomías existentes y futuras sean fácilmente convertidas en ontologías OWL. Tal proceso ayuda a identificar las similitudes y diferencias entre XBRL y OWL. Cabe mencionar que para la evaluación de sus resultados, contaron con la cooperación de analistas financieros internacionales, una empresa líder en el análisis del mercado financiero Español, y Gestifonsa[20], una firma de gestión de fondos de inversión. La traducción de taxonomías XBRL en ontologías OWL a través del proceso genérico presentado por Núñez y sus colaboradores, se limita a las empresas registradas en el CNMV, una alternativa de solución a esta limitante es la integración de las taxonomías financieras y el conjunto de datos correspondientes al modelo semántico inspirado en Linked Data propuesto en esta tesis. Esta integración fomentaría la reutilización de datos entre ambas iniciativas. Una similitud de este trabajo con el proyecto descrito por (Lara, Cantador, & Castells, 2006), es el uso ontologías OWL que resultan muy favorables en el ámbito de las finanzas y la contabilidad. Sin embargo, ninguno de ellos hace uso de Linked Data lo que se traduce en una limitada o nula capacidad para la interoperabilidad y navegabilidad entre sus datos. Además, desaprovechan los beneficios que Linked Data ofrece para la localización y conexión de datos, esto significa que también desaprovechan los beneficios de Linked Open Data cuya diferencia conceptual radica en que este último incluye la conexión de datos propios con datos externos así como la reutilización y redistribución de datos siguiendo la filosofía del Software libre que pretende que los datos sean públicos y por ende, abre un importante abanico de oportunidades para cualquier sector, actividad o negocio.

[18]Banco de España: http://www.bde.es/bde/es/
[19]CNMV: http://www.cnmv.es/portal/home.aspx
[20]Gestifonsa: http://www.cajacaminos.es/

Finalmente, el trabajo de (Núñez et al., 2008), se centra al mercado financiero Español cuyos Estados financieros se basan en el estandar de publicacion IFRS adoptado por el Ministerio de Economía de España con la publicación de la Ley 62/2003, en la que se establecen sus medidas fiscales, administrativas y sociales (Callao, Jarne, & Laínez, 2007), mientras que en este trabajo de tesis se pretende hacer uso de la norma US-GAAP ya que cuenta con una mayor aceptacion por parte de las principales empresas a nivel internacional, un ejemplo de esto se encuentra en la conslusión de (Gordon et al., 2008) en la que menciona que los Estados financieros publicados bajo US-GAAP exhiben un mayor contenido informativo en relación con IFRS.

Un enfoque similar al proceso presentado por (Núñez et al., 2008), es la construcción de modelos de información explícitos para los fondos de inversión en el mercado Español presentado por (Lara et al., 2006). Los autores presentan una taxonomía XBRL de fondos de inversión y un proceso para la traducción de taxonomías XBRL en OWL. También examinaron los beneficios relativos al uso de ontologías OWL y taxonomías XBRL para el intercambio y análisis de información de los fondos de inversión. Los autores ofrecen una alternativa de solución a los problemas de categorización e interoperabilidad de los datos financieros a través de la utilización de tecnologías semánticas. Sin embargo, su trabajo se centra en la traducción de taxonomías XBRL hacia ontologías OWL, esta traducción no es escalable respecto al creciente tamaño de los conceptos financieros y de los datos que requieren ser navegados para la obtención de información relacionada con ellos. En este trabajo de tesis, inspirándose en los principios de Linked Data (T Berners-Lee, 2009), se transforman datos XBRL en RDF para la generación de un conjunto de datos financieros cuyos datos son interoperables y se ajustan a las taxonomías financieras de Hojas de Balance, Cuenta de resultados y Estado de flujos de efectivo. Tales taxonomías son reutilizables y forman parte integral del modelo semántico inspirado en Linked Data que se presenta en este trabajo de tesis.

Entre otras iniciativas relacionadas con este trabajo de tesis, están los sistemas para el análisis de la información financiera, incluidos los sistemas de apoyo a la toma de decisiones. Algunas de estas obras se describen brevemente en la siguiente sección.

2.5.2 Trabajos para el análisis de la información financiera

Un trabajo interesante es el presentado por (Creamer & Freund, 2010), en él desarrollaron un modelo predictivo con el que demuestran como el enfoque *boosting* (Schapire, 2003), soporta dos funciones para la realización del análisis financiero, la primera

se basa en una herramienta de predicción que sirve para pronosticar el desempeño corporativo de las empresas y la segunda consiste en una herramienta interpretativa para generar árboles de decisión, que capturan la relación no lineal entre las variables contables (ratios financieros) y las variables corporativas que determinan el rendimiento de las empresas. Para su modelo predictivo, Creamer & Freund utilizan el algoritmo *AdaBoost* (Freund & Schapire, 1997), como meta-algoritmo de aprendizaje y como clasificador de las empresas que están por debajo y por encima de la media del cálculo del ratio Tobin's Q (Brainard & Tobin, 1968). Los autores realizaron experimentos de Cross-Validation (Refaeilzadeh, Tang, & Liu, 2009) a empresas de diversas regiones geográficas que cotizan en el índice bursatil S&P 500[21], empresas de Latinoamérica registradas en los ADR's[22] *(American Depositary Receipts)* y bancos domiciliados en países de América Latina. Además, comparan sus resultados con los algoritmos siguientes: *Logistic Regression* (Hosmer & Lemeshow, 1989) *Random Forest* (Breiman, 2001), y *Bagging* (Breiman, 1996). Los resultados obtenidos con *AdaBoost* indican que las grandes empresas se desempeñan mejor que las pequeñas empresas, principalmente cuando estas empresas tienen un número limitado de activos a largo plazo en relación con las ventas. Los beneficios mejoran para las grandes empresas de Latinoamérica cuando el país de residencia se caracteriza por tener un estado de derecho débil. En el caso de las compañías que cotizan en el S&P 500, el rendimiento aumenta cuando la compensación de las atribuciones es en su mayoría variable. Respecto a los estudios regionales comparativos, Creamer & Freund desarrollaron dos conjuntos de datos con poca información, el primero está conformado por 51 empresas latinoamericanas y el segundo por 104 bancos domiciliados también en Latinoamérica. Para este último identificaron un problema relacionado con la forma de integrar los datos procedentes de distintas fuentes, y en general con diferentes estándares. Los autores mencionan que este problema se encuentra implícito en este conjunto de datos y no proporcionan una solución concreta sin embargo, creen que la investigación de los mercados emergentes mejorará si se amplía el conjunto de datos, y la ejecución de los algoritmos de aprendizaje se realiza en subconjuntos agregados por regiones o sistemas de gobierno corporativo. A diferencia del modelo predictivo descrito en este párrafo, el trabajo de tesis que aquí se presenta un modelo semántico inspirado en Linked Data que proporciona las bases suficientes para generar un conjunto de datos financiero que permite la interoperabilidad entre sus datos y el análisis del estado financiero de las empresas a través de consultas basadas en SPARQL

[21]S&P 500: http://www.spindices.com/indices/equity/sp-500
[22]ADR's: http://www.investor.gov/news-alerts/investor-bulletins/investor-bulletin-american-depositary-receipts

que favorecen la realización de cálculos financieros adicionales útiles para dar soporte a la toma de decisiones.

Aplicado al contexto financiero (datos/ratios), el Análisis Envolvente de Datos *(DEA, Data Envelopment Analysis)* se utilizó para producir una medida unificada de las métricas de rendimiento (J. Zhu & Wang, 2011). Utilizando técnicas *Bootstrap* (Efron & Tibshirani, 1994), los autores proporcionan una aplicación para la evaluación del desempeño de 23 sectores manufactureros Griegos con el uso de datos financieros. Los resultados obtenidos revelan que en la primera etapa de su análisis de sensibilidad, los índices de eficiencia obtenidos estaban sesgados. Sin embargo, después de la aplicación de técnicas *Bootstrap*, el análisis de sensibilidad de los índices de eficiencia mejoró significativamente. Una aportación de este trabajo de tesis hacia la obra presentada por estos autores, radica en que el conjunto de datos financieros basado en Linked Data, les permitirá complementar su trabajo a través del suministro de ratios financieros correspondientes a las empresas registradas en esta base de conocimientos, con la finalidad de ampliar el número de empresas y las métricas que puedan ser aplicadas en el análisis DEA.

Continuando con el contexto del análisis financiero, el sistema FAST *(Fundamental Analysis Support for Financial Statements)* (Rodríguez-González et al., 2012), se basa en el uso de tecnologías semánticas que son aprovechadas a través del desarrollo de un motor de razonamiento que ofrece capacidades de inferencia basada en reglas para apoyar el proceso del análisis de inversión para un conjunto de empresas, cuya información está almacenada en formato de ontologías. El desarrollo de FAST consiste en reutilizar parte de la ontología financiera del proyecto SONAR (Gómez-Berbís et al., 2009), que favorece el almacenamiento de la información financiera necesaria para la ejecución de una parte del sistema FAST y que aporta un enfoque complementario al sistema de redes neuronales CAST (Rodríguez-González et al., 2011). Enfatizando tanto a SONAR como a CAST, el primero es un motor semántico de búsquedas financieras facultado para el rastreo y almacenamiento de información semiestructurada, así como la aplicación de *Ontology-Driven Inference* y la ejecución de estrategias para la población de ontologías mediante la aplicación de herramientas de Procesamiento de Lenguaje Natural *(NLP, Natural Language Processing)* (Maedche & Staab, 2001; Shamsfard & Barforoush, 2004). El segundo es un conjunto de soluciones para el cálculo del RSI *(Relative Strength Indicator)* utilizando técnicas de Inteligencia Artificial, estas técnicas están basadas en el uso de redes neuronales para el cálculo del RSI de manera más precisa, a lo que los autores nombran iRSI. CAST fue

utilizado en dos escenarios, el primero consistió en predecir el IBEX-35[23] del mercado de valores Español y el segundo consistió en predecir los valores para las empresas pertenecientes al IBEX-35. Los resultados son muy alentadores y ponen de manifiesto que CAST es capaz de predecir el mercado de valores Español como un conjunto y como acciones individuales pertenecientes al índice IBEX-35. El conjunto de datos financiero basado en Linked Data que se presenta en esta tesis, ofrece datos y capacidades para procesar consultas basadas en SPARQL que permiten analizar el historial de las empresas y derivar posibles predicciones, así como el cálculo de ratios financieros que apoyan el análisis fundamental financiero. Esto servirá de complemento a las funcionalidades ofrecidas por el sistema FAST.

El concepto de Sistema de Gestión de Conocimientos Financieros *(FKMS, Financial Knowledge Management System)* (Cheng, Lu, & Sheu, 2009), es un enfoque basado en ontologías y dirigido a las aplicaciones de Inteligencia de Negocios *(BI, Business Intelligence)*, específicamente para el análisis estadístico y minería de datos. FKMS tiene capacidades para la extracción, la transformación y carga de datos, la creación y recuperación de cubos de datos, el análisis estadístico y minería de datos, la gestión de experimentos con metadatos y la realización de experimentos para la obtención de resoluciones a problemas nuevos. El conocimiento resultante de cada experimento se define como un conjunto de conocimientos que consta de cadenas de datos, un modelo y parámetros utilizados. Además, los informes generados son almacenados, compartidos y difundidos. Por lo tanto, son útiles para apoyar la toma de decisiones. El enfoque de esta tesis, se basa en la generación de una base de conocimientos financieros cuyos datos son susceptibles de ser analizados y extraídos para la generación de estadísticas, análisis de empresas, cálculo de indicadores adicionales a través de consultas basadas en SPARQL, entre otras cosas que resulten de interés para los usuarios.

El uso de taxonomías y modelos de simulación en el campo de la economía y finanzas no es ajeno a la investigación, tal es el caso de Brenner & Werker, (2007) quienes ofrecen una taxonomía de los métodos de simulación existentes en esta área y muestran como los resultados obtenidos se utilizan para explicar las características económicas observadas en las empresas, examinar los sistemas económicos y predecir futuros procesos económicos. Además, proporcionan un nuevo tipo de método que ayuda a una mejor explotación de los resultados empíricos de los modelos de simulación y proponen un modelo al que llaman

[23]IBEX-35: http://www.ibex35.com/esp/aspx/Portada/Portada.aspx

modelo de simulación abductivo basándose en la inferencia por abducción (Lawson, 2002). Brenner & Werker mencionan que en el contexto de su investigación, *abducción* significa que recopilan información detallada sobre el desarrollo de diversas industrias en diferentes países con diferentes leyes de patentes, y con base a esto, clasifican los diferentes desarrollos e identifican sus fuerzas motrices subyacentes. Finalmente, argumentan que los enfoques de simulación ofrecen potencialidades científicas que aún no se explotan plenamente al utilizar la inferencia por abducción y que en particular, los modelos de simulación basados en ese tipo de inferencia tienen el potencial suficiente para ser ampliamente aceptados en la comunidad científica, porque tienden a describir, explicar y predecir mejor los procesos económicos. La iniciativa de los autores implementa un conjunto de datos en el que almacenan la informacion referente a las empresas analizadas mediante su modelo de simulación. Sin embargo, su iniciativa no contempla la interoperabilidad entre los datos que integran a su conjunto de datos, situación que puede ser aprovechada a través de la integración del conjunto de datos financiero sustentado en el modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009) propuesto en esta tesis. Esta integración permitiría ampliar las capacidades de inferencia del modelo de simulación propuesto por Brenner & Werker, (2007), complementándolas con las capacidades de inferencia propias del modelo semántico propuesto en esta investigación. Otra diferencia radica en que los autores presentan una taxonomia de los modelos de simulacion en el campo de la economía y finanzas, mientras que en esta tesis se presentan tres taxonomías financieras (Hojas de Balance, Cuenta de resultados, Estado de flujos de efectivo) que involucran los principales ratios financieros que propician el análisis fundamental de los Estados financieros de las empresas.

Un trabajo destinado hacia la búsqueda de informacion financiera es LOGIT model (Acosta-González & Fernández-Rodríguez, 2014). LOGIT model consiste en una metodología de búsqueda computacional dirigida únicamente por los datos para la selección de ratios financieros. El procedimiento utilizado en LOGIT model se basa en algoritmos genéticos que se utilizan para explorar el universo de los modelos puestos a disposición por los posibles ratios financieros existentes (con información muy redundante). El proceso de búsqueda del modelo correcto es guiado por el criterio de estimación para la dimensión de modelos descrito por (Schwarz & others, 1978). La metodología se aplica para predecir el fracaso de las empresas del sector de la construcción Español utilizando la información anual de la contabilidad pública. La construcción del modelo semántico inspirado en Linked Data de esta tesis, incluye su propio proceso de

adquisición y extracción de datos financieros para la generación de un conjunto de datos financiero basada en Linked Data a partir de documentos XBRL almacenados en el repositorio de datos EDGAR. Los datos almacenados en esta base de conocimientos, se utilizarán para consultar la información relacionada con el historial financiero de las empresas y generar posibles predicciones que ofrecen apoyo para la toma de decisiones.

El ámbito educativo también es un referente importante dentro de la investigación financiera, por ejemplo (Gomaa, Markelevich, & Shaw, 2011), desarrollaron un proyecto que utiliza herramientas interactivas de acceso a datos públicos mediante la que introducen a los estudiantes de negocios y de contabilidad en el uso del estándar XBRL. Adicionalmente, los estudiantes utilizan técnicas tradicionales para el análisis financiero mediante las que comparan los resultados de las empresas analizadas. Por lo tanto, con el proyecto se logran dos objetivos importantes, se presenta a los estudiantes los beneficios y características de XBRL y se les muestra cómo utilizar este medio para facilitar el análisis de los Estados financieros. Haciendo alusión a que el trabajo presentado por (Gomaa et al., 2011) tiene fines educativos, es importante mencionar que con la investigación a desarrollar en esta tesis, se abre la posibilidad de que los estudiantes y las personas en general interactuen con los datos que se almacenen en el conjunto de datos basado en Linked Data que aquí se presenta, con lo que se les ofrece la opción de experimentar, analizar, calcular y obtener sus propias conclusiones financieras.

El desarrollo y aplicación de una metodología para la adquisición y representación del conocimiento para desarrollar sistemas expertos en el campo del análisis financiero fue presentado por (Matsatsinis, Doumpos, & Zopounidis, 1997). Por otra parte, un motor para la búsqueda semántica de noticias financieras fue presentado por (Lupiani-Ruiz et al., 2011). Este trabajo de tesis se enfoca en los orígenes de datos semiestructurados con el fin de proporcionar datos más precisos. Las fuentes de datos no estructuradas o semiestructuradas como XBRL, suelen ser menos precisas, pero la extracción y transformación de sus datos hacia una notación estructurada mejora significativamente el análisis financiero (Matsatsinis et al., 1997), como es el caso de el modelo semántico inspirado en Linked Data que aquí se presenta. De la misma manera, las noticias financieras encontradas con el motor presentado por (Lupiani-Ruiz et al., 2011) puede ser una referencia cruzada con la información almacenada en la base de conocimientos financieros de esta tesis.

Uno de los trabajos con múltiples funcionalidades financieras es FRAANK *(Financial Reporting and Auditing Agent with Net Knowledge)*, descrito por (Bovee et al., 2005) FRAANK implementa análisis inteligente para extraer cifras contables de los Estados financieros disponibles en el repositorio EDGAR, correspondiente a la U.S. SEC. FRAANK desarrolla el razonamiento de cifras contables a través de conjuntos de etiquetas que son sinónimos de ratios financieros en una taxonomía XBRL. Como resultado, FRAANK convierte las Hojas de Balance, el estado de resultados y el Estado de flujo de efectivo en un formato XBRL-etiquetado. Esta conversión, sugiere una aproximación empírica a la evaluación y mejora de las taxonomías XBRL. Adicionalmente, FRAANK integra los datos contables con otra información financiera a disposición del público en Internet, tales como la cotización oportuna de acciones, el análisis para la previsión de ganancias, el cálculo de importantes ratios financieros y el análisis de otros indicadores financieros.

La diferencia esencial de este trabajo de tesis con respecto a FRAANK, radica en que esta se fundamenta con un modelo semántico inspirado en Linked Data que favorece la generación de una base de conocimientos financieros a partir de la extracción y transformación de los documentos almacenados en el repositorio EDGAR inspirándose en los principios de Linked Data (T Berners-Lee, 2009). La base de conocimientos financieros generada mantiene la integración e interoperabilidad de los datos y está integrada por tres taxonomías financieras (Hojas de Balance, Cuenta de resultados y Estado de flujos de efectivo) que permiten el cálculo de ratios a través de consultas basadas en SPARQL para facilitar el análisis fundamental financiero que sirva de apoyo a la toma de decisiones tanto automatizada como por parte de las personas. Además, tanto las taxonomías como la base de conocimientos financieros, están disponibles para los usuarios y permiten ser utilizadas por otros sistemas.

2.6 Conclusiones del estado del arte

El modelo semántico que se presenta en este trabajo de tesis favorece la generación de una base de conocimientos financieros ligera, escalable y reutilizable que se inspira en los principios de Linked Data (T Berners-Lee, 2009). En este sentido, la investigación que se realiza en esta tesis es una alternativa útil para servir de complemento a las iniciativas presentadas previo a estas conclusiones, proporcionando un conjunto de características nuevas y destacadas. El modelo semántico inspirado en Linked Data aborda los retos reales de la integración de datos financieros y confirma la aseveración hecha por (O'Riain, Harth,

et al., 2012), que indica la necesidad de construir un ecosistema financiero basado en el uso de los estándares Web actuales.

A diferencia del sistema FRAANK y de los trabajos relacionados previamente descritos, el modelo semántico inspirado en Linked Data que se describe en esta tesis presenta un enfoque que va más allá de la modelización Semántica de datos basados en XBRL y su carga con instancias de los datos financieros extraídos. En contraste con la típica transformación hacia RDF de documentos XBRL (García & Gil, 2010b), el modelo semántico propuesto en esta tesis proporciona versatilidad para la realización de consultas y el análisis de datos financieros, lo cual incluye la reutilización de taxonomías definidas en XBRL tales como: hoja balance, Estado de flujo de efectivo y estado de resultados. Esta reutilización no significa que se trate de un proceso para la traducción directa de taxonomías XBRL en ontologías basadas en OWL como los trabajos presentados por (Lara et al., 2006; Núñez et al., 2008) que se centran en el mercado de valores Español y que no ofrecen capacidades de escalabilidad respecto al creciente tamaño de los conceptos financieros, sino que se trata de tres taxonomías financieras (Hojas de Balance, Cuenta de resultados y Estado de flujos de efectivo) basadas en la norma US-GAAP, que forman parte integral de la base de conocimientos financieros generada a partir de un modelo semántico inspirado en Linked Data con capacidades de escalabilidad con relación al creciente tamaño de los conceptos financieros. Tanto las taxonomías financieras, como la base de conocimientos, son reutilizables por otros sistemas y sirven de alternativa de solución a los problemas de ambigüedad, inconsistencia, inexactitud, falsedad y de valores incompletos presentes en los datos financieros en línea descritos en la iniciativa de (Du & Zhou, 2012).

Con este modelo semántico inspirado en Linked Data no se pretende imitar el modelo XBRL, sino proporcionar una base de conocimientos financieros en la que se podrían analizar documentos en otros formatos que no sean XBRL y poblar la base de conocimientos a partir de ellos. La integración del conocimiento se sustenta mediante el uso de tecnologías semánticas, siendo Linked Data el principal protagonista para la publicación de un gran conjunto de datos financieros *(Financial Triple-store)* a través del procesamiento de datos XBRL para los Estados financieros publicados en formato 10-Q (U.S. SEC., 2013). Además, el modelo semántico inspirado en Linked Data contempla el uso de métodos para procesar y analizar datos financieros con el fin de hacer análisis comparativos entre las diferentes empresas, así como el cálculo de ratios financieros

adicionales que apoyen a la toma de decisiones tanto automatizada como por parte de las personas, así como servir de complemento o suministro de datos financieros para otros sistemas tales como FAST (Rodríguez-González et al., 2012), que puede hacer uso de los datos financieros almacenados en la base de conocimientos para extender las capacidades de análisis de su motor de inferencia hacia otras empresas y no limitarse a utilizar los datos de empresas registradas en las ontologías de SONAR (Gómez-Berbís et al., 2009) y por lo tanto, a las empresas comprendidas en el IBEX-35, el sistema FKMS (Cheng et al., 2009), también puede extraer los datos disponibles en la base de conocimientos, ya que le serán útiles para la creación de cubos de datos y para la aplicación de análisis estadístico y minería de datos dirigidos a la Inteligencia de Negocios, otra iniciativa que resultará beneficiada es el análisis DEA (J. Zhu & Wang, 2011), el cual puede hacer uso de los datos financieros de la base de conocimientos y medir el rendimiento de otras empresas, además de las ya utilizadas en su iniciativa. La metodología de búsqueda de información financiera LOGIT (Acosta-González & Fernández-Rodríguez, 2014) podrá incluir en sus búsquedas a la base de conocimientos de esta tesis. FRAANK (Bovee et al., 2005) es un proyecto muy completo que no utiliza Linked Data pero que factiblemente puede complementar sus funcionalidades mediante la explotación de los datos almacenados en la base de conocimientos basada en Linked Data que se presenta en esta tesis. Cambiando el contexto hacia el ámbito educativo, en el trabajo de (Gomaa et al., 2011) los alumnos y maestros de contabilidad y finanzas se beneficiarán con la base de conocimientos que aquí se presenta, teniendo la opción de experimentar, analizar, calcular y obtener sus propias conclusiones financieras.

Para concluir el estado del arte y dicho lo anterior, son múltiples los beneficios a obtener por la generación de una base de conocimientos financieros cimentada por un modelo semántico inspirado en Linked Data que permite la navegación entre los conceptos de la base de conocimientos, proporcionando flexibilidad para el usuario final permitiéndole principalmente, explorar los datos y hacer sus propias conclusiones o descubrimientos financieros.

Capítulo 3

Planteamiento de las hipótesis y proceso de validación

Resumen. Los problemas que se presentan durante una investigación para la creación de sistemas que permitan la generación de conocimiento y el soporte a la toma de decisiones en ámbitos financieros, no son problemas que tengan una solución trivial ni sencilla. Desde el punto de vista de este trabajo de tesis, estos sistemas requieren de un modelo semántico que sirva de soporte para su correcta funcionalidad y de una base de conocimientos común que ayude en las miles de decisiones financieras que se llevan a cabo diariamente en todo el mundo. Durante la investigación para el cumplimiento de estos propósitos, surgen ciertas teorías o hipótesis que requieren de un proceso de validación organizado. De acuerdo con lo expresado, en este capítulo se plantean las hipótesis propias de esta investigación, así como el proceso necesario para su validación.

3. Planteamiento de las hipótesis a resolver

Como se ha presentado en la sección del estado del arte, existen varias iniciativas que se han investigado y desarrollado a lo largo de los años para dar solución a diversos problemas financieros. A diferencia de estos, en esta tesis se pretende la generación de un modelo semántico inspirado en Linked Data, lo que implica una investigación que a su vez involucra una serie de procedimientos de carácter técnico-científico que permiten el aislamiento, categorización y clasificación de aquellos elementos económico-financieros provenientes de distintas fuentes de información disponibles en la Web, para transformarlos en datos financieros que serán explorables, leídos e interpretados (computables) por un ordenador con el fin de generar conocimiento financiero, mejorar la calidad estructural de los datos, enlazar los datos con fuentes de información externa y apoyar a la toma de decisiones tanto automatizada como por parte de las personas. Estas pretensiones requieren de un proceso de investigación que de manera inherente conlleva al surgimiento de ciertas hipótesis, las cuales provienen del nexo entre la teoría y la realidad empírica de la presente tesis doctoral. En tal sentido, una hipótesis sirve para orientar y delimitar una investigación, dándole una dirección definitiva a la búsqueda de la solución de un problema (Tamayo, 2004).

Considerando las pretensiones descritas previamente para este trabajo de tesis, se procede a la descripción de las hipótesis a verificar:

- **Hipótesis *H1*:** el modelo semántico inspirado en los principios de Linked Data propuesto, permite poblar una base de conocimientos financieros a partir de la integración de fuentes de datos externas.
- **Hipótesis *H2*:** los datos procesados en la base de conocimientos financieros a partir del modelo semántico propuesto, permiten la reutilización de datos con terceros a través de Linked Data.
- **Hipótesis *H3*:** utilizar tecnologías semánticas para transformar y organizar la información financiera publicada en la Web ayuda a mejorar la calidad estructural de los datos.
- **Hipótesis *H4*:** una base de conocimientos financieros fundamentada en Linked Data, favorece el análisis el análisis fundamental financiero para apoyar la toma de decisiones tanto automatizada como por parte de las personas.

Uno de los propósitos de las hipótesis es servir de directrices en una investigación (Tamayo, 2004), y en consecuencia, resulta imperativo llevar a cabo su evaluación o validación para ser comprobadas. De acuerdo con lo expresado y por la naturaleza de esta investigación, es primordial la realización de una propuesta de validación organizada que incluya el desarrollo de las aplicaciones de Software necesarias para la comprobación de las hipótesis planteadas en esta tesis. Dicha propuesta de validación es descrita en el apartado siguiente.

3.1 Propuesta de validación de las hipótesis

Para la validación de las hipótesis *H1, H2, H3 y H4,* es necesario apoyarse de las tecnologías semánticas que propicien el desarrollo de las aplicaciones de Software necesarias para facilitar la comprobación de cada una de la hipótesis planteadas. Con base en las diferentes etapas de los ciclos de vida descritos en el Estado del arte (véase Sección 2.4.14) y subrayando que las diferentes fases del ciclo de vida de Linked Data no existen en forma aislada o iteran en una secuencia estricta (Auer et al., 2013), se diseñó un proceso de extracción y transformación de datos hacia RDF considerando principalmente a los documentos basados en el estándar XBRL para la generación de una base de conocimientos financieros inspirada en Linked Data. Esencialmente, este proceso de extracción y transformación de datos comprende las seis fases que se muestran en la Figura 11 y que se describen a manera de introducción en esta sección, pero que se profundizan a detalle en la sección siguiente.

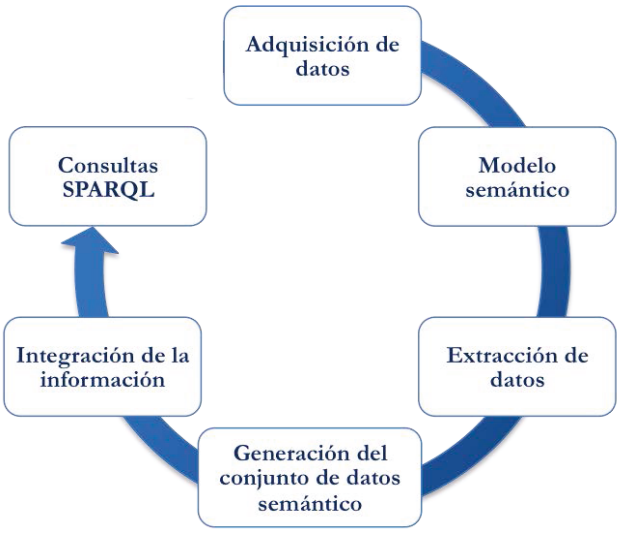


Figura 11. Proceso para la generación de la base de conocimientos financieros inspirada en Linked Data

La fase de *Adquisición de datos* se centra en la obtención de los Estados financieros trimestrales publicados bajo el estándar XBRL por parte de las empresas y bajo la norma

US-GAAP, el *Modelo semántico* es el núcleo para la creación del modelo de datos que sirve de base para la transformación a RDF de los datos obtenidos en la fase de *Adquisición de datos* y que se efectúa en la fase de *Extracción de datos,* la *Generación del conjunto de datos semántico* serializa los datos transformados en la fase de *Extracción de datos,* la fase de *Integración de la información* busca la vinculación de los datos almacenados en el conjunto de datos semántico con otras fuentes de datos. Finalmente, la fase de *Consultas SPARQL* permite la adquisición de información por parte de los usuarios.

Es importante resaltar que las aplicaciones de Software a desarrollar son un medio tecnológico para proceder a la validación de las hipótesis planteadas, por lo tanto, el modelo semántico inspirado en Linked Data, la población de la base de conocimientos financieros inspirada en Linked Data, la calidad estructural de los datos financieros obtenidos, los cálculos financieros para el análisis fundamental a partir de consultas basadas en SPARQL, el apoyo a la toma de decisiones tanto manual como automatizada, y la reutilización de datos con terceros a través de Linked Data, serán validados cualitativa y cuantitativamente por un grupo de expertos en finanzas y contabilidad con el objetivo de evaluar las hipótesis planteadas.

3.2 Proceso de transformación de datos financieros para la validación de las hipótesis planteadas

Uno de los funcionamientos de la Web Semántica se basa en poner a disponibilidad grandes cantidades de datos RDF, no como islas de datos incomunicadas, sino como una red de bases de datos interconectadas. Hasta la fecha, este requisito no se ha cumplido ampliamente, lo que lleva a amplías críticas porque obstaculizan la tarea y el progreso de los desarrolladores que desean crear aplicaciones para la Web Semántica. Sin embargo, gracias al movimiento *Open Data*, existe una valiosa oportunidad para rectificar en parte esta situación haciendo conjuntos de datos libres a partir de fuentes de datos ya existentes tales como Wikipedia, Musicbrainz, Geonames, Wordnet y DBLP, entre otras, haciéndolos disponibles como datos RDF interconectados a gran escala (Bizer et al., 2007).

Con el fin de poyar el movimiento *Open Data,* el proceso de extracción e interconexión de datos financieros para la validación de las hipótesis planteadas en la presente tesis, involucra la transformación a tripletas RDF de los datos almacenados en los documentos XBRL para generar un conjunto de datos semántico inspirado en los principios de Linked Data, explorable y computable que comprende seis fases principales (véase Figura 11) en

las que: a) los datos financieros son extraídos y procesados; b) el modelo semántico es aplicado; y c) la infraestructura de consultas es cargada con datos semánticos. El proceso de transformación es iterativo y cada fase permite ser refinada en múltiples iteraciones. La explicación de cada una de estas fases y su papel en todo el proceso de transformación, así como el apoyo que estas brindan para dar solución a cada una de las hipótesis planteadas, es proporcionada de manera detallada a continuación:

- **Adquisición de datos:** esta fase tiene la particularidad de ser el punto de partida para la generación de la base de conocimientos financieros mediante la que se llevará acabo la validación de cada una de las hipótesis planteadas en esta tesis doctoral. A través de un *Crawler*, se rastrean e identifican las fuentes de información financiera con Estados financieros publicados trimestralmente, es decir, bajo el formato 10-Q del repositorio EDGAR (U.S. SEC., 2013), siguiendo el estándar XBRL y bajo la norma US-GAAP. Tales Estados financieros son descargados y almacenados para su posterior proceso de transformación en tripletas RDF. Considerando que el proceso de publicación de los Estados financieros por parte de las empresas no es precisamente continuo, lo que significa que estas no publican su información financiera en periodos de fechas obligatoriamente exactos, es necesario repetir esta fase manualmente cuantas veces sea necesario para identificar y adquirir los Estados financieros más recientes, con la finalidad conservar la base de conocimientos con información actualizada.
- **Modelo semántico:** se trata de la fase central del proceso de transformación en general, es en donde se crea el modelo de datos que sirve de apoyo para la validación de las cuatro hipótesis planteadas. En esta fase, se conceptualizan dos modelos semánticos, el primero es un modelo Mixto que integra características del diseño canónico de datos y del modelo Entidad-Atributo-Valor. El segundo, es un modelo alternativo al modelo Mixto y se basa en el modelo Entidad-Atributo-Valor con reificación. Esta fase es análoga a la de diseño de esquemas de bases de datos para los Sistemas de Gestión de Base de Datos Relacionales (SGBDR) o RDBMS *(Relational Database Management System)*. Sin embargo, el modelo semántico va más allá del tipo de lógica procedimental *(procedural logic)* (Georgeff, Lansky, & Bessiere, 1985) que subyace al algebra relacional, en la que a través de un modelo (modelo relacional) se define un conjunto de operaciones (procedimiento) que paso a paso computan una respuesta sobre las relaciones (Bonner & Kifer, 1994). A diferencia de este tipo de lógica, el modelo semántico se centra en la definición de relaciones

semánticas entre conceptos y clases conceptuales, que en el contexto de esta tesis, captura todos los aspectos de la información financiera expresada en los documentos XBRL. Adicionalmente, proporciona capacidades para la realización de tareas analíticas con Estados financieros cruzados, así como el procesamiento de consultas versátiles basadas en SPARQL. Aunque el modelo semántico se basa en la estructura de los Estados financieros XBRL, no se trata de una traducción literal de todas las taxonomías que conforman su estructura (véase Figura 1), pero sí reutiliza algunas de sus taxonomías financieras tales como Hojas de Balance, Estado de flujo de efectivo y Estado de cuenta de resultados. Otro aspecto importante del modelo semántico es la reutilización de vocabularios como *Time Ontology* para la representación de los periodos de publicación de los Estados financieros y de *Payments Ontology*, que permite representar la información correspondiente a los gastos de una organización en formato de Linked Data.

- **Extracción de datos:** en esta fase, se extraen los datos almacenados en los Estados financieros XBRL obtenidos en la fase de Adquisición de datos. Estos Estados financieros son Hojas de Balance, Estados de cuentas de resultados, Flujos de efectivo e información adicional que resulta relevante. Los datos financieros se extraen junto con las taxonomías XBRL y sus bases de referencia *(Linkbases)* (XBRL-España, 2006). Entre los datos que se extraen se incluyen etiquetas, referencias, definiciones y cálculos. Adicionalmente, para cada estado financiero se extraen metadatos complementarios tales como fechas, detalles de las empresas y los periodos de presentación de los Estados financieros. En esta fase es importante asegurarse que los datos se extraen correctamente y que la transformación a tripletas RDF no contiene información irrelevante, con el objetivo de ser almacenadas en un conjunto de datos semántico y de servir de apoyo a la validación de la hipótesis *H2* .
- **Generación del conjunto de datos semántico:** esta fase está estrechamente relacionada con la fase anterior y consiste en serializar los datos en forma semántica según la definición del Modelo semántico. La serialización se realiza utilizando el marco de trabajo Apache Jena (Lindörfer, 2010) y el almacenamiento de los datos financieros en formato de tripletas RDF se efectúa dentro del repositorio semántico Virtuoso Open-Source (Erling & Mikhailov, 2009). El conjunto de datos permite ser consultado a través del SPARQL endpoint proporcionado por el repositorio Virtuoso Open-Source y cuyo enlace es ofrecido en el capítulo de validación del

presente trabajo de tesis. Inspirarse en los principios de Linked Data favorece la integración y navegación de los datos financieros extraídos en la fase anterior, lo que significa que los conceptos poseen identificadores únicos definidos como URI's que pueden ser fácilmente desreferenciados a través de consultas basadas en SPARQL con la finalidad de recuperar más información. Como ya se indicó, la generación del conjunto de datos se realizó de acuerdo con el Modelo semántico, siguiendo su esquema, sus relaciones y sus definiciones ontológicas, lo que favorece la validación de las hipótesis *H3* y *H4*.

- **Integración de la información:** después de que el conjunto de datos semánticos es cargado con los datos extraídos y transformados en tripletas RDF a partir de los Estados financieros XBRL, su forma semántica se reduce significativamente a las barreras de la integración de datos con otros conjuntos de datos. En el dominio de Linked Data este proceso se denomina interconexión de datos *(Data interlinking)* (Michael Hausenblas, 2009). Aquí es donde los conceptos del conjunto de datos local están relacionados con los conceptos de fuentes de datos externas, apoyando a la validación de la hipótesis *H4*. Por ejemplo, los conjuntos de datos públicos disponibles en la LOD cloud, los datos del World Bank Linked Data (Worldbank, 2012) o con los datos internos de alguna empresa (Allemang, 2010). Para la validación de esta hipótesis, en el Capítulo 6 se procede a la realización de los experimentos que permitan el descubrimiento de enlaces en la LOD cloud a través de DBPedia y cuya información se encuentre relacionada con las empresas almacenadas en el conjunto de datos generado. El motivo por el que estos experimentos se centran en DBPedia, radica en que DBPedia es el núcleo de la LOD cloud y por lo tanto, permite acceder y navegar a través de otros conjuntos de datos (Auer et al., 2007; Bizer et al., 2009). Para complementar los experimentos para el descubrimiento de enlaces en la LOD cloud, se realizan experimentos de *Precision and Recall* (Ting, 2010) de manera manual, con la finalidad de identificar a partir de los enlaces descubiertos, ¿Cuáles enlaces realmente contienen información relacionada con las empresas almacenadas en el conjunto de datos generado? Como ya se ha señalado, los detalles de estos experimentos se describen en el Capítulo 6.
- **Consultas SPARQL:** se trata de la fase final del proceso para la generación de la base de conocimientos financieros. En esta fase, los datos ya han sido extraídos y vinculados entre sí, y están listos para fines analíticos. Esto se traduce en apoyo para la validación de las hipótesis *H3* y *H4* respectivamente, porque incluye el

cálculo de ratios financieros e indicadores adicionales que favorecen el análisis fundamental de las empresas y a la toma de decisiones tanto automatizada como por parte de las personas. Las consultas basadas en SPARQL se utilizan para proporcionar meta-modelos de los indicadores calculados y para proporcionar un marco de trabajo que favorezca su reutilización por parte de aplicaciones externas que requieran procesar el cálculo de datos financieros para su beneficio y con base en el modelo semántico inspirado en Linked Data.

Como se menciona en este capítulo, el núcleo del proceso de extracción e interconexión de datos financieros es el modelo semántico inspirado en Linked Data descrito en la segunda fase de dicho proceso, por tal motivo, es importante proporcionar su descripción más a detalle, lo cual ocurre en el capítulo siguiente.

Capítulo 4

Modelos semánticos inspirados en Linked Data para la publicación de datos financieros

Resumen. Reutilizar los datos financieros ubicados en entornos distribuidos en la Web para organizarlos, estructurarlos y vincularlos inspirándose en los principios de Linked Data (T Berners-Lee, 2009), es un proceso que desde la perspectiva de esta tesis, requiere de un modelo semántico de datos que sirva de soporte para la generación de una base de conocimientos financieros. Siguiendo este contexto, en el presente capítulo se presentan dos modelos semánticos de datos que permitirán la generación de una base de conocimientos financieros inspirada en Linked Data. El primer modelo integra características de diseño canónico y del modelo Entidad-Atributo-Valor, está orientado hacia la normalización, el cálculo y la búsqueda de datos financieros. El segundo modelo, se sustenta en el modelo Entidad-Atributo-Valor con reificación y con características de búsqueda y navegación entre los datos.

4. Modelado semántico de datos

Hoy en día, la información es generada por los datos ubicados en un entorno distribuido pero vinculado. Por esta razón es necesario el uso de modelos comunes para la representación de la información y para ponerla a disposición de terceros. En este sentido, los modelos semánticos juegan un papel crucial para la representación de la información relativa a un dominio dado (Gangemi, 2005). Gracias al modelado semántico, es posible representar la información basada en vocabularios comunes y compartidos, lo que permite a los sistemas basados en el conocimientos encontrar, recuperar y tomar decisiones basadas en ellos. En este contexto, los lenguajes más utilizados para la representación de modelos semánticos de datos, son RDF(S) y OWL (basado en RDF). Aunado a estos, SPARQL es el lenguaje de consulta de modelos basados en RDF recomendado por el consorcio de la W3C (Dimitrov, 2012; W3C, 2013). Técnicamente, dichos lenguajes son primordiales para generar la base de conocimientos financieros que se fundamenta en los modelos semánticos que se describen en el presente capítulo.

4.1 Modelo Mixto de datos financieros

Considerando el preámbulo del párrafo previo y aludiendo a los fundamentos teóricos relacionados con el Modelo conceptual, los Modelos de datos y los Modelos de datos semánticos descritos previamente en el Estado del Arte, se procede con la descripción conceptual del modelo semántico Mixto inspirado en los principios de Linked Data (T Berners-Lee, 2009) que sustenta la investigación que se presenta en este trabajo de tesis y, que como ya se ha mencionado, es donde se crea el modelo de datos enfocándose en la definición entre conceptos y clases conceptuales capturando todos los aspectos de la información financiera expresada en los documentos XBRL.

La Figura 12, es una representación conceptual del modelo semántico inspirado en Linked Data que no incluye la interconexión de datos con otras fuentes de información. Se trata de una representación mixta que surge a partir de la integración entre el modelo Entidad-Atributo-Valor (EAV) y el diseño canónico, cuyo propósito es el de proporcionar un modelo semántico normalizado que favorezca el cálculo de indicadores financieros y la navegabilidad de los datos. Cada uno de los componentes que integran a este modelo son descritos a continuación (Radzimski, Sánchez-Cervantes, Garcia-Crespo, & Temiño-Aguirre, 2014):

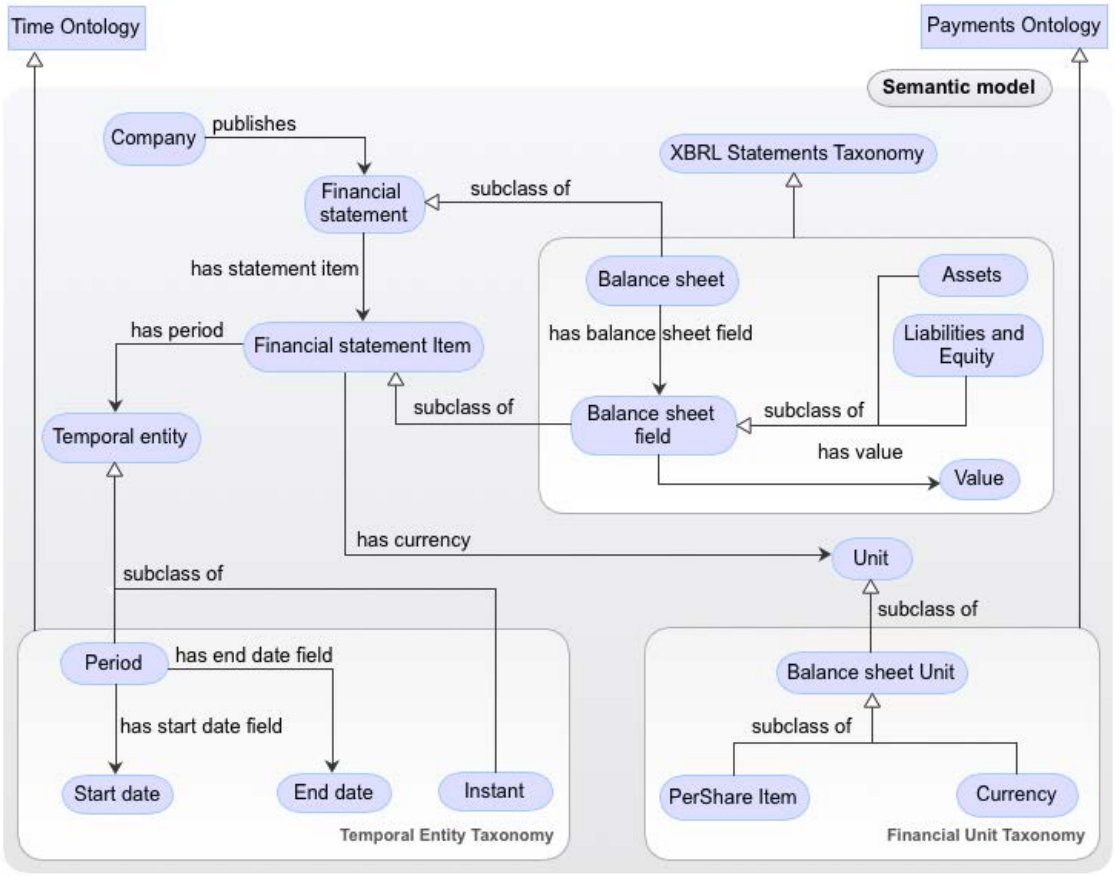


Figura 12. Modelo semántico Mixto de datos financieros inspirado en Linked Data

- **Empresa *(Company)*:** dentro del modelo semántico, la Empresa representa a la organización que publica sus Estados financieros en la Web. En este caso, se considera hacer uso del repositorio EDGAR de la U.S. SEC.
- **Estado Financiero *(Financial Statement)*:** los Estados financieros son documentos publicados por las empresas. Entre los documentos publicados se encuentran Hojas de balance, Cuenta de estado de resultados y Estado de flujo de efectivo. Cada uno de estos documentos contiene información contable de la situación de la empresa en un momento muy preciso, usualmente al final de un ejercicio anual, un semestre o de un trimestre.
- **Ítem del Estado Financiero *(Financial Statement Ítem)*:** dentro del modelo semántico, este componente se relaciona con el Estado financiero y representa a uno de los tres Estados financieros publicados por la empresa. Esto significa que un Ítem de Estado Financiero es la representación (instancia) de alguno de los siguientes Estados financieros Hojas de balance, Cuenta de estado de resultados y Estado de flujo de efectivo.

- **Entidad Temporal *(Temporal Entity)*:** como ya se mencionó, los Estados financieros son publicados cada determinado periodo de tiempo por parte de las empresas. Las fechas y el instante en el que estos son publicados, se representan por la Entidad Temporal, que se relaciona con el Ítem de Estado financiero, es decir, la instancia de las Hojas de balance, Cuenta de estado de resultados y Estado de flujo de efectivo.
- **Ontología de Tiempo *(Time Ontology)*:** esta ontología proporciona un vocabulario para expresar propiedades acerca de las relaciones topológicas entre instantes e intervalos de tiempo, esto significa que permite representar información referente a fechas y horas (Hobbs & Pan, 2004, 2006). En el contexto del modelo semántico Mixto, el vocabulario *Time Ontology* sirve de apoyo para la publicación en formato Linked Data de los periodos en los que los Estados financieros han sido divulgados.
- **Taxonomía de la Entidad Temporal *(Temporal Entity Taxonomy)*:** esta taxonomía reutiliza el vocabulario *Time Ontology* y proporciona los metadatos correspondientes al periodo de fechas y el instante en el que se publicaron los Estados financieros. Esta taxonomía y se integra de los componentes siguientes:
    - **Periodo *(Period)*:** corresponde al espacio de tiempo en el que se publicaron los Estados financieros.
    - **Fecha de inicio *(Start date)*:** indica el mes, día y año en el que inicia el periodo de publicación de los Estados financieros.
    - **Fecha de finalización *(Start date)*:** corresponde al mes, día y año en el que finaliza el periodo de publicación de los Estados financieros.
    - **Instante *(Instant)*:** indica el instante en el que fueron publicados los Estados financieros, lo que significa que son válidos para una fecha o fecha y hora específicas.
- **Unidad *(Unit)*:** los Ítems de los Estados Financieros *(Financial Statement Ítems)*, o lo que es lo mismo, las Hojas de balance, la Cuenta de estado de resultados y el Estado de flujo de efectivo, tienen un tipo de moneda asignado. La Unidad representa toda la información concerniente a las unidades de medida utilizadas en los Estados financieros entre las que destacan los Dólares, acciones, Euros o Dólares por acción. En este aspecto, la asignación de las unidades de medida en los Estados financieros XBRL, se realiza de a cuerdo al conjunto de unidades estándar establecidas en el Registro de Unidades *(Units Registry)*. El objetivo del Registro de

Unidades, es ser un conjunto de datos público en línea en el que se documenten estas unidades y su uso (Pryde et al., 2013). Es importante mencionar que las unidades de medida no se limitan a los tipos de moneda utilizados en los Estados financieros, también se manejan unidades de longitud, masa, tiempo y temperatura, entre otras (XBRL-International, 2013).

- **Ontología de Pagos *(Payments Ontology)*:** se trata de una ontología que ofrece un vocabulario de uso general para representar la información de los gastos de una organización, no es específica de algún determinado gobierno o de solicitudes realizadas por gobiernos locales. Dentro del modelo semántico, el vocabulario de *Payments Ontology* es reutilizado en la *Financial Unit Taxonomy* porque permite que los datos correspondientes a los gastos de las organizaciones puedan ser representados en formato de Linked Data, tal vez de forma experimental o como parte del desarrollo de herramientas y procesos reutilizables por terceros (Reynolds, 2011).
- **Taxonomía de Unidad Financiera *(Financial Unit Taxonomy)*:** en los documentos XBRL, todos los valores tienen alguna unidad de medida asignada (XBRL-España, 2006). Esta taxonomía, suministra los metadatos de las unidades de medida aplicadas a los Estados financieros, y se integra de los siguientes componentes:
    - **Unidad de medida de la Hoja de balance *(Balance sheet Unit)*:** dentro de la conceptualización del modelo semántico, la Unidad de medida de la Hoja de balance hereda de Unidad *(Unit)* la información correspondiente al tipo de moneda asignado a los valores de este estado financiero.
    - **Ítem por Acción *(PerShare Ítem)*:** se relaciona con la Unidad de medida de la Hoja de balance y es utilizado para los conceptos que se miden en Dólares por acción. La referencia de la Unidad *(Unit)* en el documento de instancia, que en este caso es la Hoja de balance, requiere la mención de la moneda en el numerador y de la cuota en el denominador. De esta manera, el valor final de la medición en la Unidad de medida se mostrará como USD por acción *(USD per share)* o Rupia por acción *(Rupee per share)*, entre otros. Un ejemplo de esto son los Beneficios por acción *(Earnings per share)*, que es son indicadores utilizados en el análisis de la Hoja balance para medir la rentabilidad por acción de una empresa en un periodo de tiempo determinado (Dodge, 1991).

- **Moneda *(Currency)*:** también se relaciona con la Unidad de medida de la Hoja de balance y se utiliza para todos los conceptos financieros que denotan una cantidad representada en una moneda. Esta Unidad de medida se rige por la Estándar Internacional para los códigos de Moneda *(International Standard for Currency Codes) Currency Codes* - ISO 4217 (ISO, 2008a). Dos ejemplos de la representación de datos bajo este estándar son el Dólar estadounidense y el Franco Suizo. El primero se representa como USD en donde US corresponde al código de país según el estándar ISO-3166 (ISO, 1997) y la D es de Dólar. Y el segundo, se representa por CHF, donde CH es el código de Suiza para la norma ISO-3166 y la F es de Franco.

- **Taxonomías XBRL de los Estados Financieros *(XBRL Statements Taxonomy)*:** este componente representa la reutilización de las taxonomías correspondientes a los Estados financieros. Estas taxonomías se estructuran por XML Schemas y *linkbases* que definen la estructura de cada estado financiero mediante los *linkbases* siguientes: etiquetas, referencias, presentación, cálculo y definición. Todos ellos previamente descritos en la Sección 2.3.3 del Estado del Arte del presente trabajo de tesis.
- **Ejemplo de la Taxonomía de Hoja de balance *(Balance sheet example)*:** dentro del modelo semántico y como su nombre lo indica, este componente representa un ejemplo de la taxonomía de Hoja de balance basada en la norma US-GAAP (véase Figura 13), que contiene los ratios financieros que servirán de base para el proceso de extracción y transformación en RDF de los datos contenidos en las Hojas de balance XBRL. La representación conceptual del proceso de transformación de estas Hojas de balance en RDF incluye los siguientes componentes:
  - **Hoja de balance *(Balance sheet)*:** es una instancia del Estado financiero *(Financial Statement)* y representa a una Hoja de balance XBRL, cuyos ratios y valores son transformados en tripletas RDF dentro de un proceso de extracción y transformación.
  - **Campo de la Hoja de balance *(Balance sheet field)*:** es una instancia del Ítem del Estado Financiero *(Financial Statement Ítem),* y representa a todos los ratios financieros contenidos en la Hoja de balance *(Balance sheet)*. Es importante resaltar que los ratios de una Hoja de balance se organizan en

grupos por ejemplo, Activos, Activos no corrientes, Pasivos y Pasivos no corrientes.

- **Activos *(Assets)*:** en el modelo semántico Mixto, un Activo es una instancia del Campo de la Hoja de balance *(Financial Statement field)*, que corresponde al grupo de los Activos dentro de las Hojas de balance.
- **Pasivos y Patrimonio *(Liabilities and Equity)*:** al igual que los Activos, un Pasivo representa la instancia del Campo de la Hoja de balance *(Financial Statement field)*, perteneciente al grupo de Pasivos en una Hoja de balance *(Balance sheet)*.
- **Valor *(Value)*:** representa el valor asignado a los Campos de las Hojas de balance *(Balance sheet field)*. Ofreciendo una descripción más detallada, este componente representa el valor asignado a los ratios que integran la Hoja de balance. En el modelo semántico, se conceptualiza el valor asignado a los Activos *(Assets)* y a los Pasivos y Patrimonio *(Liabilities and Equity)*.

La integración del diseño canónico y el modelo EAV, proporciona un modelo semántico que se ajusta a las taxonomías de los Estados financieros basados en la norma US-GAAP que son descritas en la sección siguiente.

4.2 Taxonomías de los modelos semánticos de datos financieros

El modelo semántico Mixto conserva los principios básicos del diseño canónico o CDM (véase Sección 2.4.7.2) (Dell & Dell, 2010) incluyendo sus atributos, asociaciones y su semántica, lo que favorece a la definición de entidades correspondientes al dominio financiero y a la integración de taxonomías financieras representadas mediante RDF(S), además de proporcionar la flexibilidad, eficiencia y capacidades para la realización de consultas centradas en las entidades (Nombre de la empresa), que son características propias del modelo EAV (Anhøj, 2003; Nadkarni et al., 1999). La Figura 13, muestra el grafo de la Taxonomía financiera de Hoja de balance basada en la norma US-GAAP. Complementariamente, de la Figura 13.1 a la Figura 13.8, se muestran los principales ratios financieros (subclases) que conforman a dicha taxonomía. Por otra parte, las Figuras 14 y 15, muestran de manera general las taxonomías de Estado de flujo de efectivo y de Cuenta de estado de resultados.

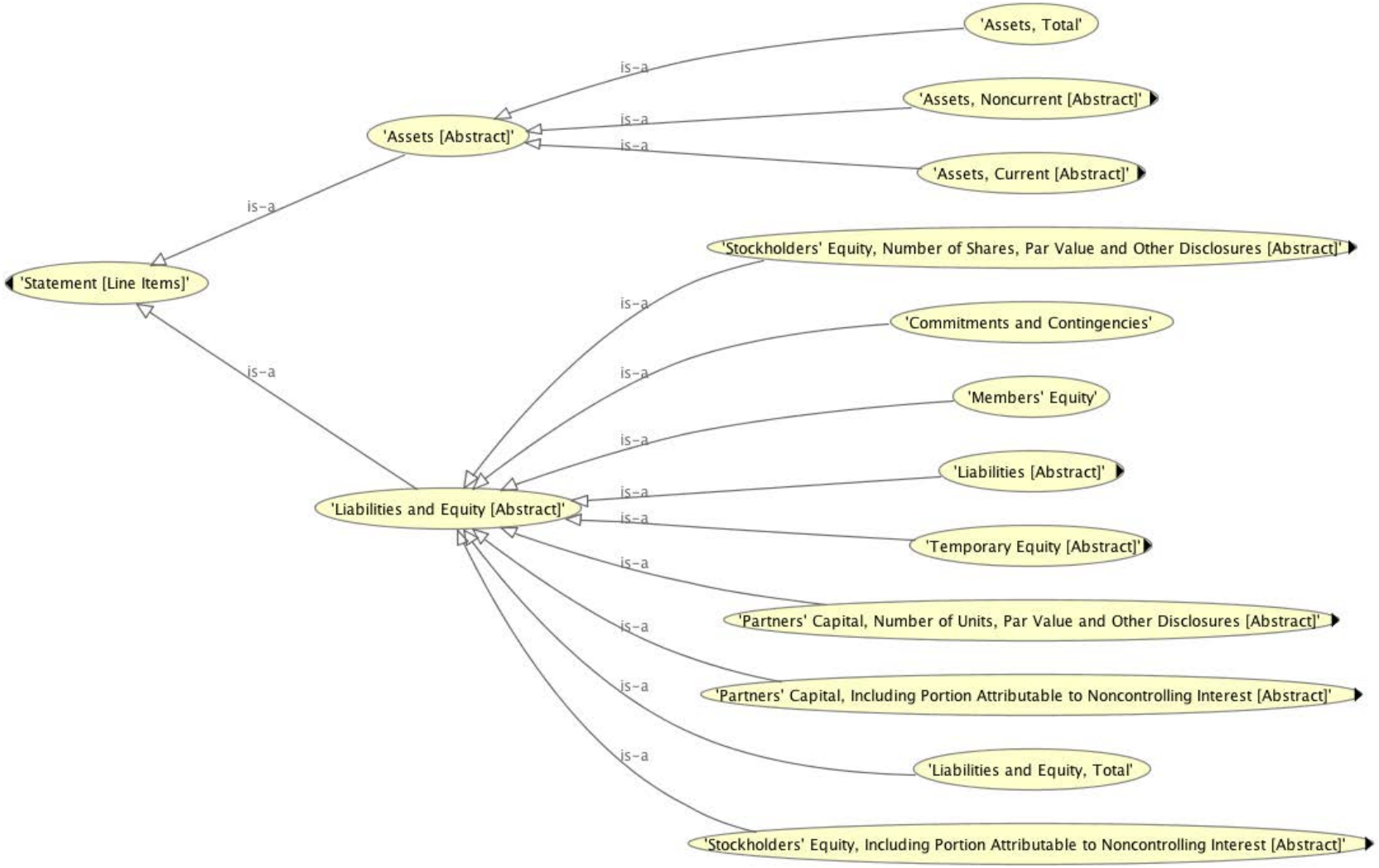


Figura 13. Taxonomía RDF(S) de Hoja de balance *(Balance sheet)* basada en la norma US-GAAP

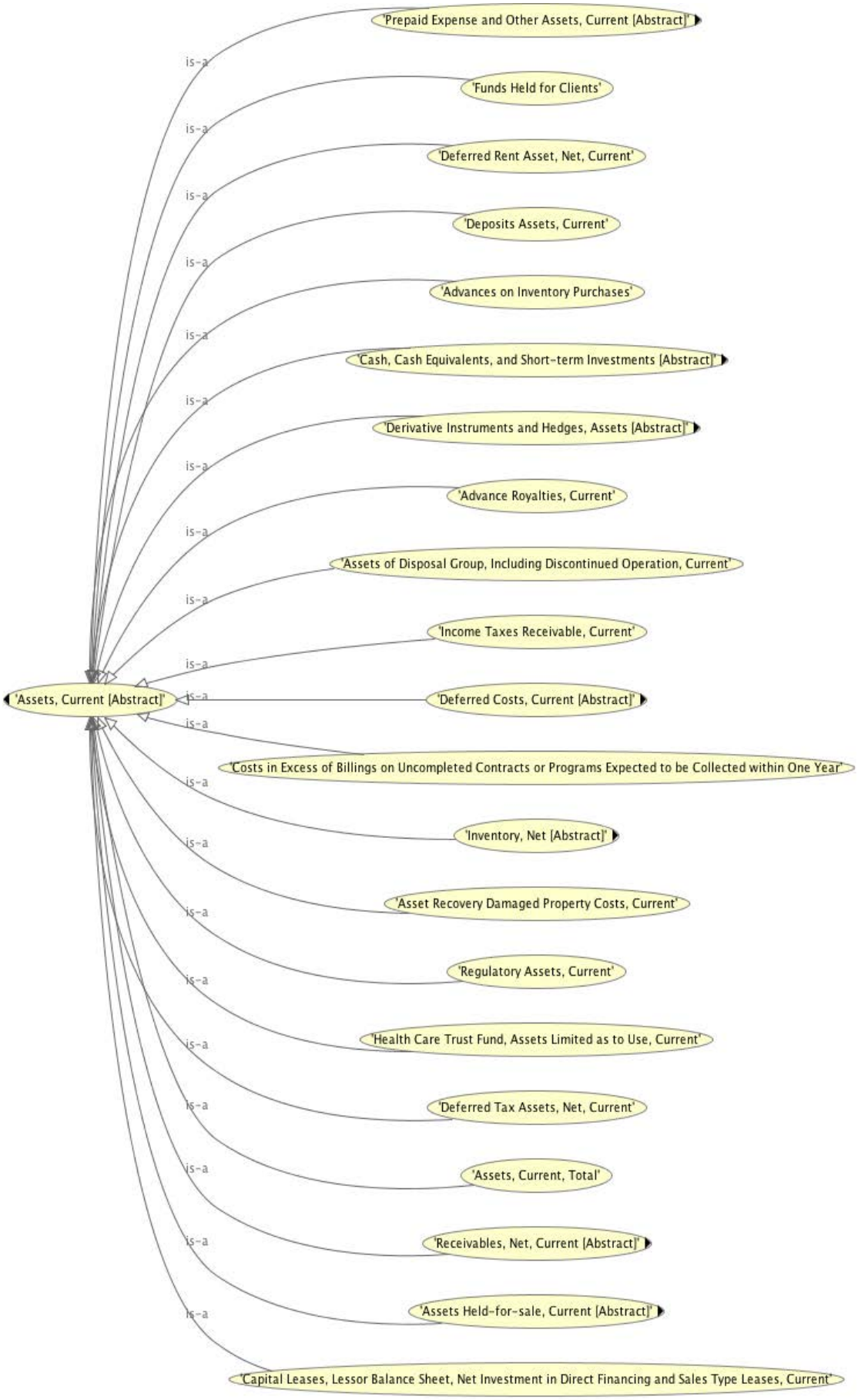


Figura 13.1 Ratios (Subclases) del Activo circulante *(Current Assets)* de la taxonomía de Hoja de balance

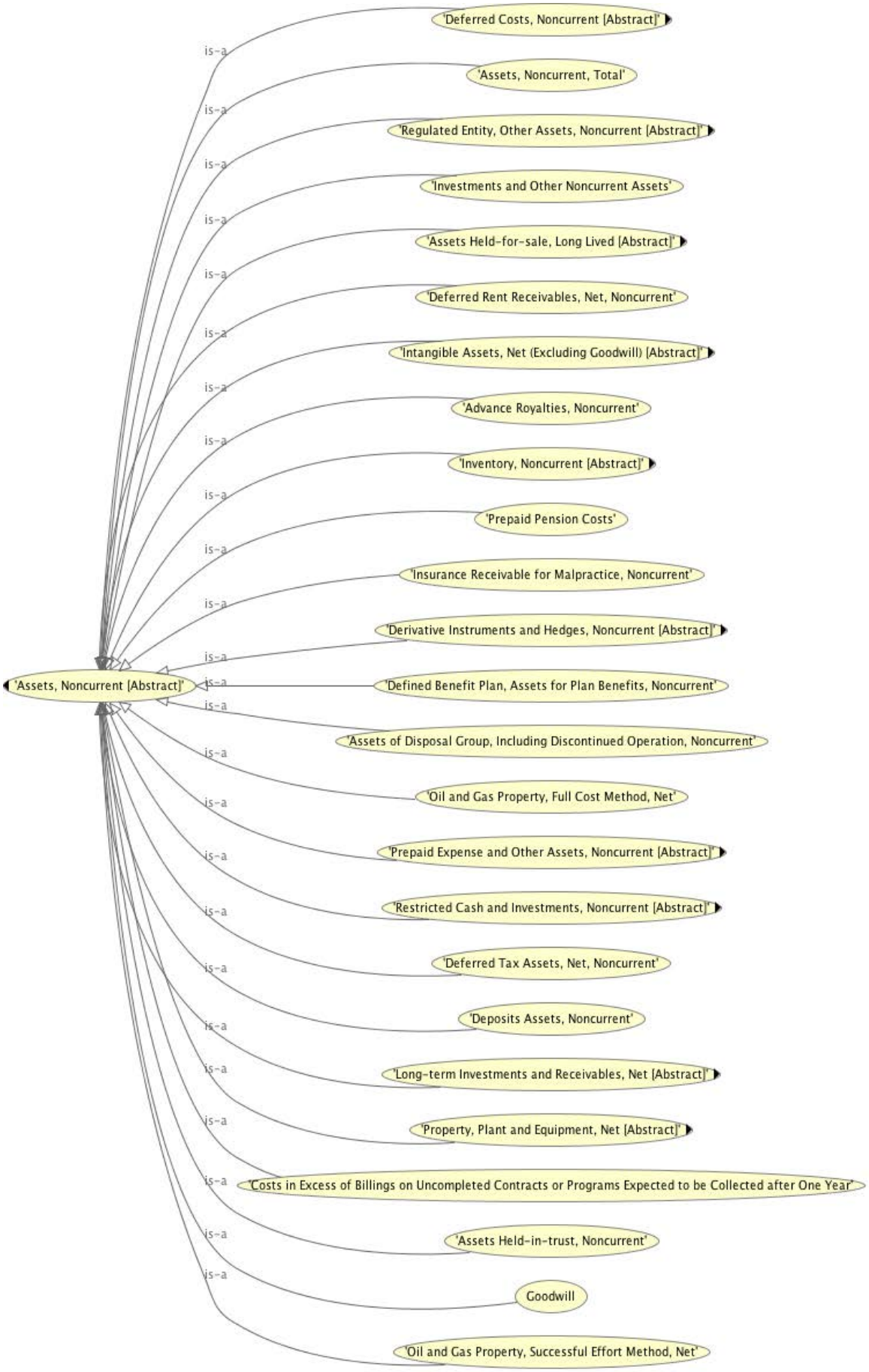


Figura 13.2 Ratios (Subclases) del Activo fijo *(Current Assets)* de la taxonomía de Hoja de balance

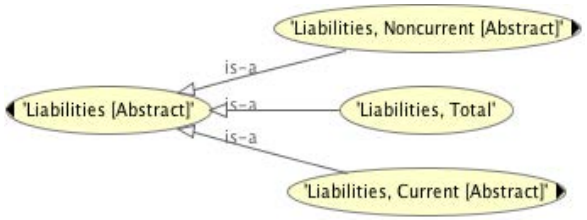


Figura 13.3 Ratios (Subclases) de los Pasivos *(Liabilities)* de la taxonomía de Hoja de balance

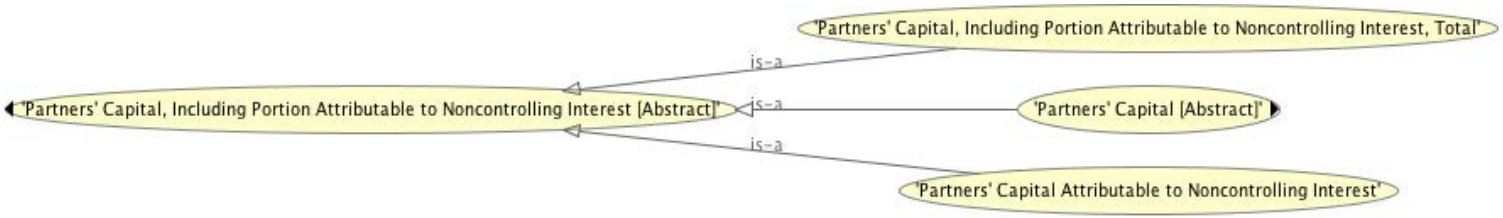


Figura 13.4 Ratios (Subclases) del Capital de socios incluyendo la parte atribuible a la participación no controladora *(Partners' Capital, Including Portion Attributable to Noncontrolling Interest)* de la taxonomía de Hoja de balance

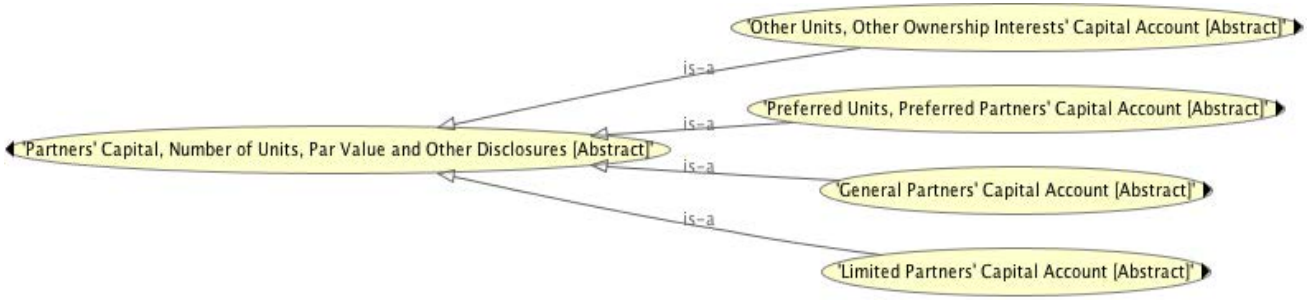


Figura 13.5 Ratios (Subclases) del Capital de socios, número de unidades, valor nominal y otras cláusulas *(Partners' Capital, Number of Units, Par Value and Other Disclosures)* de la taxonomía de Hoja de balance

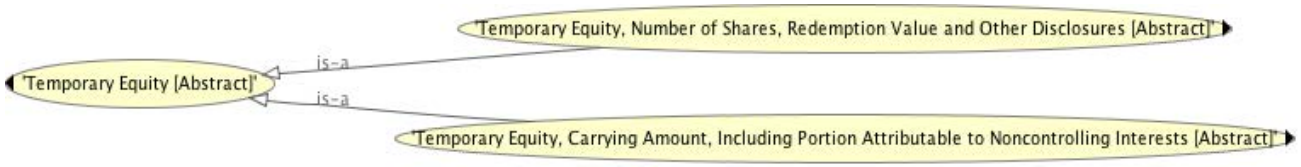


Figura 13.6 Ratios (Subclases) de la Equidad temporal *(Temporary Equity)* de la taxonomía de Hoja de balance

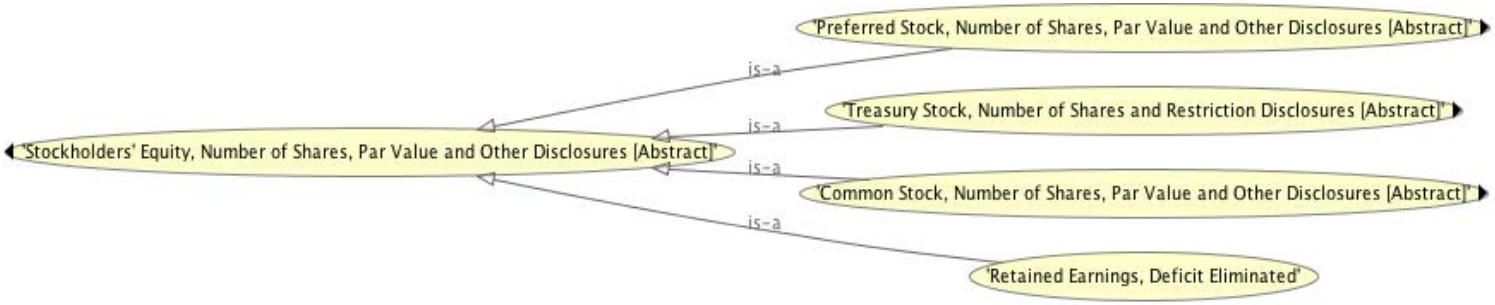


Figura 13.7 Ratios (Subclases) del Capital contable, número de acciones, valor nominal y otras cláusulas *(Stockholders' Equity, Number of Shares, Par Value and Other Disclosures)* de la taxonomía de Hoja de balance

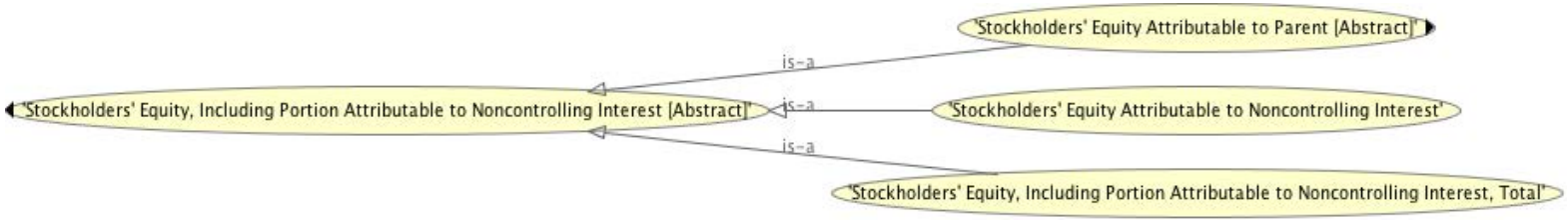


Figura 13.8 Ratios (Subclases) del Capital contable, incluyendo la porción atribuible a la participación no controladora *(Stockholders' Equity, Including Portion Attributable to Noncontrolling Interest)* de la taxonomía de Hoja de balance

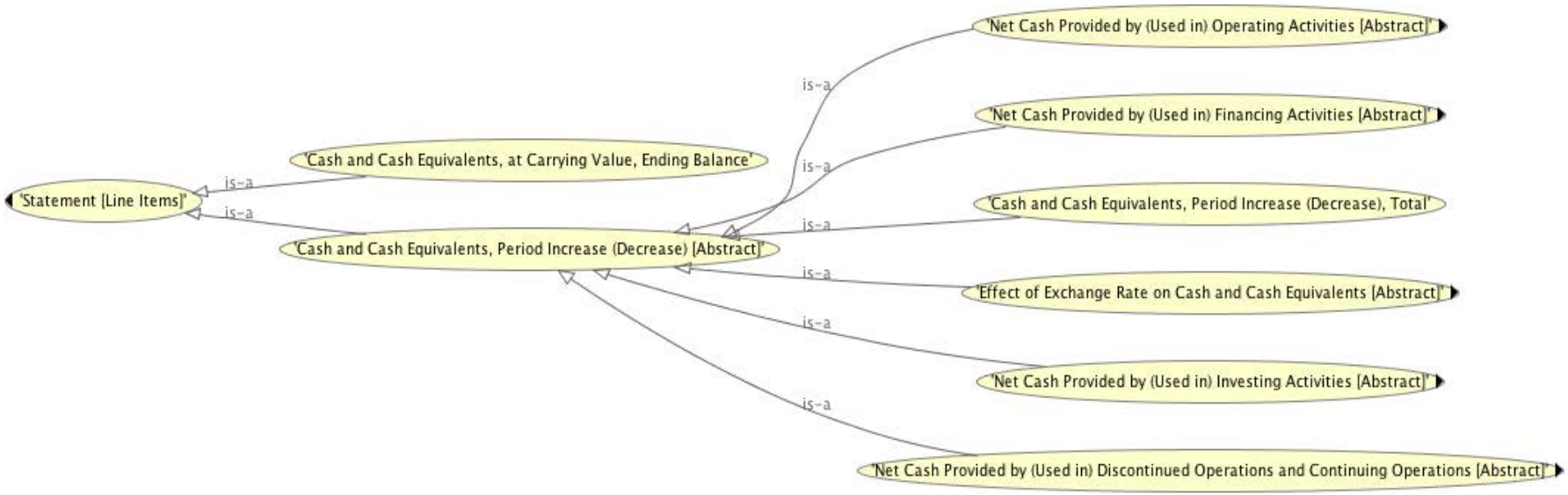


Figura 14. Taxonomía RDF(S) de Estado de flujo de efectivo *(Cash flow)* basada en la norma US-GAAP

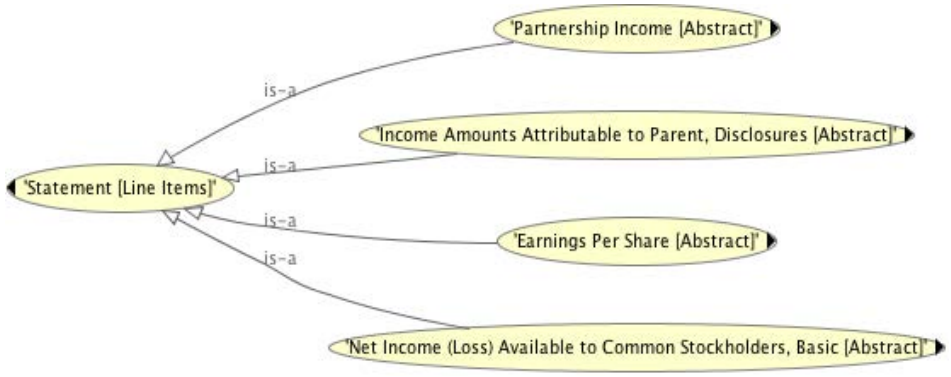


Figura 15. Taxonomía RDF(S) de la Cuenta de estado de resultados *(Income statement)* basada en la norma US-GAA

Cada una de las taxonomías financieras presentadas en las Figuras previas, forman parte integral de la base de conocimientos que se genera a partir de los modelos semánticos inspirados en Linked Data que se describen en el presente Capítulo. Sin embargo, para la validación de las hipótesis planteadas en esta tesis doctoral, sólo se hace uso de la taxonomía de las Hoja de balance (véase Figura 13). El motivo principal del uso de este estado financiero radica en que las Hojas de balance reflejan la situación financiera de una empresa en un tiempo determinado (Penman, 2009), lo que resulta muy conveniente ya que las Hojas de balance utilizadas para la transformación en RDF, son almacenadas y publicadas trimestralmente (Formato 10-Q) bajo el estándar XBRL en el repositorio EDGAR de la U.S SEC. (U.S. SEC., 2013). Por lo que no es objetivo de esta tesis, procesar cálculos ajenos a los indicados para el análisis fundamental financiero. Sin embargo, tanto las taxonomías financieras, como la base de conocimiento inspirada en Linked Data son reutilizables para el uso que más convenga a los usuarios.

La taxonomía de Hoja de balance facilita el desarrollo de cálculos que permiten la obtención de los indicadores financieros especificados en las Tablas 1, 2 y 3 del Estado del Arte, y cuyo propósito es el de facilitar el análisis contable de las empresas para el apoyo de la toma de decisiones tanto automatizada como por parte de las personas. La Tabla 5, contiene una breve descripción de los principales ratios de la taxonomía de Hoja de balance, que ayudan a cumplir este propósito.

RATIOS DE LA TAXONOMÍA DE LA HOJA DE BALANCE PARA EL ANÁLISIS CONTABLE DE LAS EMPRESAS	
RATIO	**DESCRIPCIÓN**
Activos, *(Assets)*	Parte de la Hoja de balance que refleja, en unidades monetarias, todas las partidas o cuentas donde figuran los bienes, inversiones, derechos de cobro y otros recursos que la empresa posee y de los que espera obtener beneficios económicos.
Activo circulante, *(Current Assets)*	Parte del activo que es líquido o puede convertirse fácilmente en efectivo. Es sinónimo de "Activo corriente o Circulante".
Efectivo y sus equivalentes, *(Cash And Cash Items)*	Dinero neto y todos sus equivalentes susceptibles a ser convertidos en dinero en efectivo.
Valores negociables, *(Marketable Securities)*	Títulos representativos de derechos que pueden transmitirse entre titulares.
Cuentas y documentos por cobrar, *(Accounts And Notes Receivable)*	Es un derecho de cobro sobre clientes producto de una transacción comercial. Puede estar documentada con un pagare o con un efecto cambiario.
Provisiones para cuentas de cobros dudosos o incobrables, *(Allowances for Doubtful accounts)*	Comúnmente se denominan Provisión para insolvencias. Se realizada para cubrir el importe de las deudas de dudoso cobro o ante la posibilidad de que los clientes vayan a ocasionar una insolvencia por falta de pago. Cuando las cuentas a cobrar se consideran definitivamente incobrables se aplica la provisión dotada en su momento contra la cuenta de resultados y si la provisión no alcanza para liquidar la deuda por la diferencia se lleva contra resultados como incobrable. El concepto previsión, es diferente a las provisiones, es utilizado por las compañías para prever futuros pagos como por ejemplo, la liquidación futura de los impuestos de sociedades o pago de dividendos a accionistas.
Ingresos no derivados del trabajo, *(Unearned Income)*	Son los conocidos como Ingresos atípicos o ajenos a la explotación. Ingreso extraordinario, no relacionado con la actividad normal de la sociedad (Por ejemplo, la venta de un inmueble, las plusvalías generadas por la cartera de valores, entre otros). Es sinónimo de "Ingreso extraordinario".
Inventarios, (Inventories)	Es la relación Activos (existencias) para ser vendidos en el curso normal de la operación en forma de materiales o suministros, y también para ser consumidos por la compañía en el proceso de producción. Los inventarios también recogen los bienes comprados y almacenados para su elaboración, transformación en productos terminados. Así mismo los terrenos u otras propiedades de inversión que se tienen para ser vendidos a terceros. En el caso de un prestador de servicios, los inventarios incluirán el coste de los servicios para los que la compañía aún no haya reconocido el ingreso de la operación correspondiente.

Tabla 5. Ratios de la taxonomía de Hoja de balance (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

RATIOS DE LA TAXONOMÍA DE LA HOJA DE BALANCE PARA EL ANÁLISIS CONTABLE DE LAS EMPRESAS	
RATIO	**DESCRIPCIÓN**
Gastos pagados por adelantado, *(Prepaid Expenses)*	Se conocen como "Gastos anticipado" y forman parte del activo de la Hoja de balance como periodificación contable. Se trata de recoger una salida de efectivo, pago por adelantado, de bienes y servicios que se recibirá en un futuro próximo.
Otros activos circulantes, *(Other current assets)*	Es la parte del activo puede convertirse fácilmente en efectivo. Es sinónimo de "Activo corriente o circulante".
Total activos corrientes, *(Total current assets)*	Es el conjunto de Activos que la empresa posee y que están ligados al ciclo normal de explotación (no superior a un año), y aquellos cuyo vencimiento, enaApache Jenación o realización no se espera que se produzca en el corto plazo (no superior a un año).
Activo fijo, *(Non-Current Assets)*	Es la parte del activo no destinada a la venta sino a permanecer en un principio en la sociedad de manera indefinida: las instalaciones, planta de producción y oficinas por mencionar algunos.
Otras inversiones, *(Other Investments)*	Son activos a disposición de la compañía normalmente no incluidos dentro de los denominados Activos corrientes.
Propiedad, planta y equipo, *(Property Plant and Equipment)*	Es el activo fijo fundamental de la compañía que es necesario para la actividad empresarial. El valor de la propiedad, planta y equipo, entre otros. Puede estar valorado en el balance de la compañía, como Activo fijo no corriente, por el coste histórico, valor razonable o valor neto de mercado por aplicación de la amortización contable. Suelen ser amortizados durante la vida estimada del activo, de acuerdo a unos coeficientes fiscales, para preservar su valor de reposición a largo plazo.
Depreciación acumulada, *(Accumulated Depreciation)*	Amortización sistemática del valor de un activo a lo largo de su vida útil, con el fin de recoger el valor del bien a efectos de su reposición al final de la vida útil del mismo.
Activos intangibles, *(Intangible Assets)*	Son activos de carácter no monetario, sin apariencia física, que resulta identificable. Forman parte de este tipo de activos los derechos, patentes, licencias u otros activos inmateriales cuyo valor es a veces difícil de cuantificar.
Amortización de activos intangibles, *(Amortization of Intangible Assets)*	En un contexto contable, la dotación a la amortización sistemática del valor por el que se va depreciando un activo a lo largo de su vida útil, por su uso, por el transcurso del tiempo, por haber cumplido con su fin o por otros motivos de esta índole.
Otros activos, *(Other Assets)*	Es la parte de los activos de menor importancia, tales como la periodificación e inmateriales.

Tabla 5 (Continuación). Ratios de la taxonomía de Hoja de balance (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

RATIOS DE LA TAXONOMÍA DE LA HOJA DE BALANCE PARA EL ANÁLISIS CONTABLE DE LAS EMPRESAS	
RATIO	**DESCRIPCIÓN**
Total activos, *(Total Assets)*	En el balance de situación se reflejan, el total de unidades monetarias que se corresponden con los bienes, inversiones de todo tipo, derechos y otros recursos que la empresa posee. Es lo que tiene la compañía.
Pasivo circular, *(Current Liabilities)*	Se denomina también Pasivo circulante o pasivo exigible a corto plazo. Son las obligaciones de pago, tales como dividendos diferidos, el crédito comercial, y los impuestos no pagados, que surjan en el curso normal de un negocio y de vencimiento del pago inferior a un año.
Cuentas y documentos por pagar, *(Accounts and Notes Payable)*	Es un compromiso de pago en fechas futuras, que se clasifican como corrientes si son por un plazo inferior a un año y si son no corrientes a un plazo superior al anterior.
Otros pasivos corrientes, *(Other Current Liabilities)*	Son las obligaciones que una compañía que espera liquidar en el transcurso del ciclo normal de explotación; obligaciones cuyo vencimiento o extinción se espera que se produzca a corto plazo (un año como máximo a partir de la fecha de cierre del ejercicio), en particular las obligaciones para las cuales la empresa no disponga de un derecho incondicional a diferir su pago en dicho plazo.
Total pasivos corrientes, *(Total Current Liabilities)*	Es la suma de las cuentas por pagar de una empresa que esta ligado al ciclo normal de explotación, habitualmente un año.
Pasivos no corrientes, *(Non-Current Liabilities)*	Son las deudas de la compañía recogidas en el pasivo no clasificables como pasivo corriente y por tanto a largo plazo. Se conoce como Pasivo fijo.
Bonos, hipotecas y otras deudas a largo plazo, *(Bonds Mortgages and Other Long Term Debt)*	Los Bonos son títulos de renta fija con un vencimiento a más corto plazo que las obligaciones, que las compañías emiten para financiar determinadas actividades a un plazo superior al ciclo de explotación. El importe en el momento de su emisión el bonista (Inversor que compra bonos o acciones) lo recupera al vencimiento y por el que percibirá un interés normalmente fijado y explicito. Pueden ser valores emitidos al portador y negociables en los mercados de valores. La hipoteca es una garantía real que vincula la propiedad de un bien inmueble al cumplimiento de una obligación de pago. En caso de impago, el acreedor tiene derecho a ejecutar la hipoteca para resarcirse de la deuda contraída.
Compromisos y pasivos contingentes, *(Commitments and Contingent Liabilities)*	Se refiere a operaciones realizadas por una compañía por la que se garantizan las obligaciones contraídas por un tercero, la posibilidad de pagar determinadas cantidades que dependen de acontecimientos futuros. Un pasivo contingente sería una demanda pendiente.

Tabla 5 (Continuación). Ratios de la taxonomía de Hoja de balance (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

RATIOS DE LA TAXONOMÍA DE LA HOJA DE BALANCE PARA EL ANÁLISIS CONTABLE DE LAS EMPRESAS	
RATIO	**DESCRIPCIÓN**
Total pasivos no corrientes, *(Non-Redeemable Preferred Stock)*	Elementos del pasivo no clasificables como pasivo corriente.
Capital contable, *(Stockholder's Equity)*	Se conoce como Patrimonio neto, su valor es el de los activos minorados por los pasivos de la compañía. Por tanto lo componen las aportaciones realizadas por sus socios o partícipes, en forma de capital social, así como los beneficios no distribuidos tanto en forma de reservas en todas sus consideraciones así como la autofinanciación. Está compuesto por los fondos propios, los ajustes por cambio de valor, y las subvenciones, previsiones, donaciones recibidas.
Acciones preferidas rescatables, *(Redeemable Preferred Stocks)*	Son las acciones preferentes redimibles, se refiere a un tipo de acciones que está sujeto a ser devuelto a la organización que emite a partir de una fecha específica a un determinado precio. Este tipo de acciones no es exigible para los primeros años después de que la acción se emite. Si las tasas de interés caen por debajo de la tasa de emisión de la acción, la empresa puede recomprar las acciones a los inversores y emitir nuevas acciones a un tipo de interés más bajo.
Acciones preferidas no redimibles, *(Non-Redeemable Preferred Stock)*	Es un tipo de acciones preferentes que no pueden ser devueltas a la empresa emisora. Las acciones preferentes se refiere a acciones propiedad de una compañía que tiene una demanda más alta de los activos y las ganancias de las acciones ordinarias. Los dividendos sobre acciones preferentes se pagan antes que la de las acciones comunes, así como en el caso de liquidación de activos en una quiebra. No tienen derecho a voto para los inversores, y goza de ciertos derechos. Éstos pueden ser la prioridad en caso de liquidación de la sociedad o el cobro de dividendos.
Acciones comunes, *(Common Stocks)*	Título que representa una parte alícuota del capital social de una sociedad anónima o una comanditaria por acciones. Puede estar o no desembolsada, ser nominativa o al portador. Tiene un valor nominal, otro contable y otro de mercado, siendo este último su cotización si está admitida su negociación en un mercado organizado de valores.
Otro capital contable, *(Other Stockholder's Equity)*	Se explicita en la Hoja de balance de la compañía para distinguirlo de la partida de Capital social, representa la aportación efectuada por los accionistas de la empresa. Hay varios conceptos que definen al Capital contable como son, el derecho de los propietarios sobre los activos netos que surgen por aportaciones de los dueños, las transacciones y otros eventos o circunstancias que afectan a una entidad y el que se ejerce mediante reembolso o distribución. Representa el patrimonio de los accionistas integrado pos sus aportaciones de capital realizadas por encima del valor nominal de las acciones. Representa todos los recursos de los que dispone una entidad para realizar de sus operaciones y que han sido aportados por fuentes internas de la entidad (dueños o propietarios, socios o accionistas), por lo que surge la obligación de la entidad de retribuirles, ya sea en efectivo, bienes, servicios, derechos, o por un interés residual en la empresa.

Tabla 5 (Continuación). Ratios de la taxonomía de Hoja de balance (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010)

4.3 Interconexión del modelo Mixto de datos financieros

Al ampliar la conceptualización del modelo semántico Mixto de datos financieros inspirado en Linked Data, se obtienen dos capas adicionales al propio modelo, que no forman parte directamente de su funcionalidad pero que permiten la interconexión de datos *(Data Interlinking)* con fuentes de datos externas, por ejemplo, DBpedia que está disponible a través de la LOD cloud.

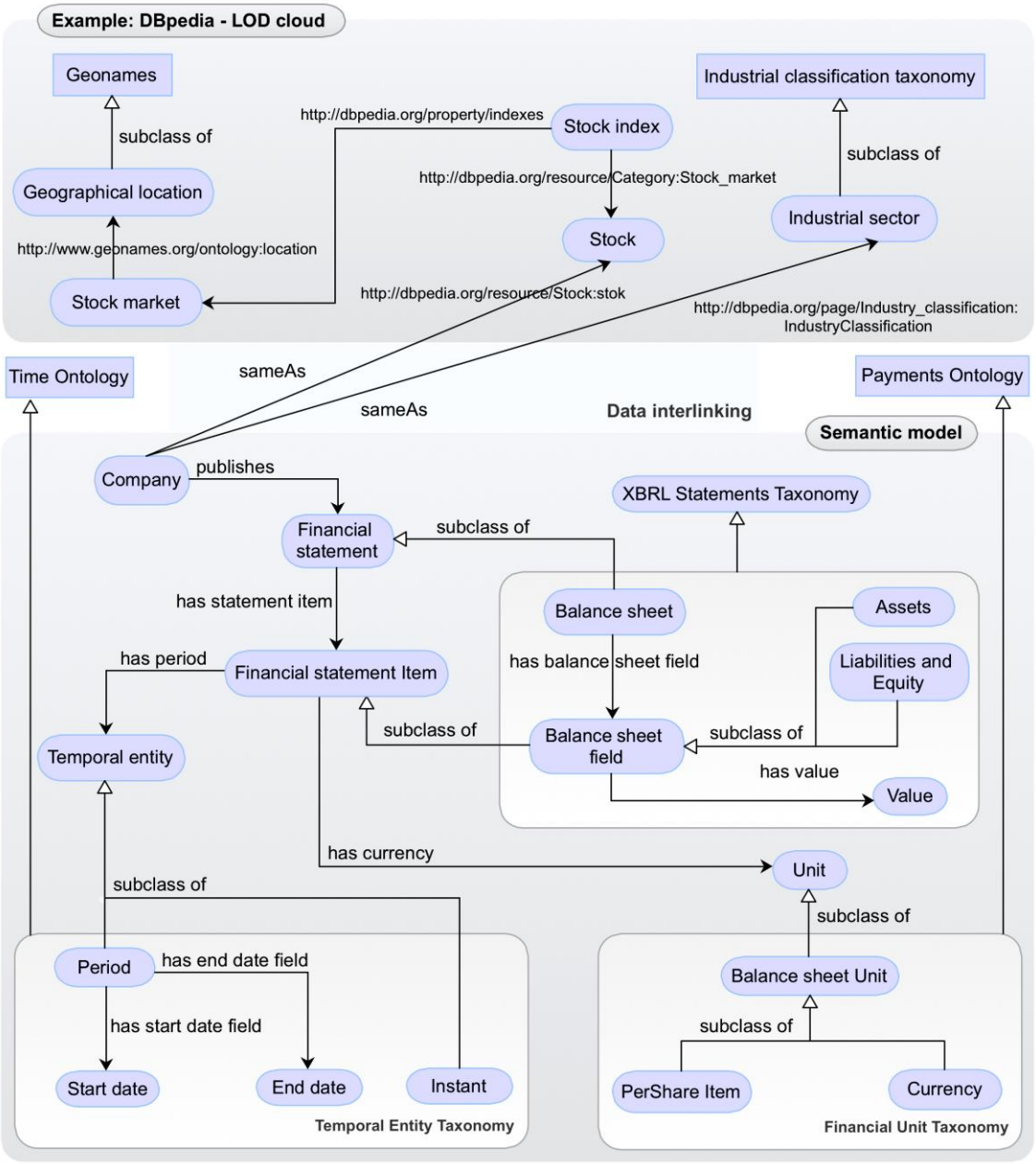


Figura 16. Modelo semántico Mixto de datos financieros inspirado en Linked Data e interconexión de datos

La Figura 16, muestra el modelo semántico con la integración de la capa de Interconexión de datos *(Data interlinking)* y del Conjuntos de datos externos utilizando

como ejemplo a DBpedia. Estas capas se integran a través de relaciones definidas por los URIs descritos en los ficheros RDF y que son utilizados sobre el protocolo HTTP para asegurarse de que cualquier recurso puede ser buscado y accedido en la Web de datos, sin olvidar que los URIs no sólo son direcciones, sino que son los identificadores de los recursos disponibles en esta Web, lo que permite que los datos sean consultados (desreferenciados) mediante consultas basadas en SPARQL y que se conecten con otras fuentes de información.

Dentro de las capas que se incluyen en el modelo semántico, y a nivel de relaciones entre las clases, subclases, valores y taxonomías que lo integran, se proporciona una explicación de cada uno de estos elementos, con la finalidad de identificar la función que tiene cada uno de ellos (Radzimski et al., 2014):

- **Modelo semántico:** esta capa se compone de varias clases y subclases que integran la funcionalidad de la base de conocimientos financieros que se genera a partir de este Modelo semántico. Para su descripción, esta capa se clasifica en dos grupos, Clases y Subclases. El primer grupo incluye a *"Company",* esta clase contiene información de las empresas y enfocándose en el modelo EAV (Nadkarni et al., 1999), *"Company"* es la entidad sobre la que se centran las búsquedas basadas en SPARQL. Además, se relaciona con *"Stock"* e *"Industrial sector"*, correspondientes al conjunto de datos externos DBpedia, núcleo de la LOD cloud (Auer et al., 2007; Bizer et al., 2009). Cada determinados meses, las empresas publican sus Estados financieros *"Financial statement"*, tales como Hoja de balance, Estado de flujo de efectivo y Cuenta de estado de resultados, que se relacionan con la Clase *"Financial Statement Item"* y que se representan de manera general en la *"XBRL Statements Taxonomy"*. Aunque conceptualmente, en el modelo semántico esta taxonomía es representada como una sola, en realidad representa a las taxonomías de cada Estado financiero publicado por las empresas y es reutilizada en el modelo semántico porque contiene los XML Schema y los *linkbases* con la información de estos Estados financieros (véase Sección 2.3.3 del Estado del Arte). Las clase *"Financial Statement Item"* contiene valores, por lo que se relaciona con la clase *"Value"* y se complementa con los metadatos aportados por su relación con las clases *"Temporal entity"* y *"Unit"*, que son súper clases de *"Period"*, *"Instant"* y *"Balance sheet Unit"* las cuales corresponden al segundo grupo de esta explicación. En este mismo contexto, la clase *"Balance sheet field",* como subclase de *"Financial statement*

Item", se beneficia de todas las propiedades, atributos y relaciones que esta última tiene con otras clases. Las subclases *"Period"* e *"Instant"*, contienen los valores del instante y la fecha en la que se publicaron los Estados financieros como documentos XBRL. Por otro lado, las subclases *"Balance sheet Unit"*, *"PerShare Item"* y *"Currency"*, contienen datos sobre las unidades financieras aplicadas en los Estados financieros. Dentro del modelo semántico, se incluye el ejemplo de una implementación de la taxonomía de Hoja de balance presentada en la Figura 13, como subclase de *"XBRL Statements Taxonomy"*. Esta implementación incluye las clases y subclases que son necesarias para la publicación de los ratios financieros correspondientes a la Hoja de balance de alguna empresa, basándose en el modelo semántico inspirado en Linked Data que aquí se describe.

- **Conjuntos de datos externos:** esta capa se representa como un ejemplo de los conjuntos de datos externos relacionados a DBpedia y que incluyen a los siguientes dominios de información: a) Localización geográfica: como el vocabulario Geonames[24] que permite obtener información geográfica de los mercados de valores, incluyendo países, ciudades y códigos postales, entre otros; b) Stock Index: *(Market Index Stock)*[25], tales como el Standard & Poor's 500, FTSE 100, Dow Jones o el Nasdaq, por mencionar algunos. En los que se negocian las acciones y bonos de empresas; c) Stock: este indicador compara el precio de cierre de una acción con su escala de precios durante un período determinado de tiempo; y d) Sector Industrial *(Industrial sector)*: este vocabulario proporciona la información necesaria para la clasificación de las empresas en función de su rama industrial. Cada uno de estos componentes aumenta el potencial y la funcionalidad de la base de conocimientos financieros por la interconexión de sus datos con los datos contenidos en la LOD cloud a través de la propiedad *owl:sameAs*, permitiendo la búsqueda y navegación de datos financieros y no-financieros entre ambos espacios de datos.
- **Interconexión de datos *(Data Interlinking)*:** la interconexión de datos facilita la conexión y navegación entre los datos financieros y no-financieros publicados en la base de conocimientos generada y los datos contenidos en fuentes de datos externas como DBpedia y otros conjuntos de datos que forman parte de la LOD cloud.

[24]Geonames: http://DBpedia.org/page/GeoNames
[25]Stock Marked: http://DBpedia.org/page/Category:Stock_market

Dentro del modelo semántico inspirado en Linked Data que se describe en el presente capítulo, la transformación de datos financieros hacia RDF utiliza los ratios definidos en la taxonomía de Hoja de balance basada en la norma US-GAAP como una jerarquía de clases, lo que representa la inherente naturaleza que caracteriza a los datos semánticos, así como su estructura y el cómo sus diferentes partes (clases, subclases y valores) se relacionan unas con otras.

4.4 Trancisión de la conceptualización del modelo semántico hacia la base de conocimientos financieros inspirada en Linked Data

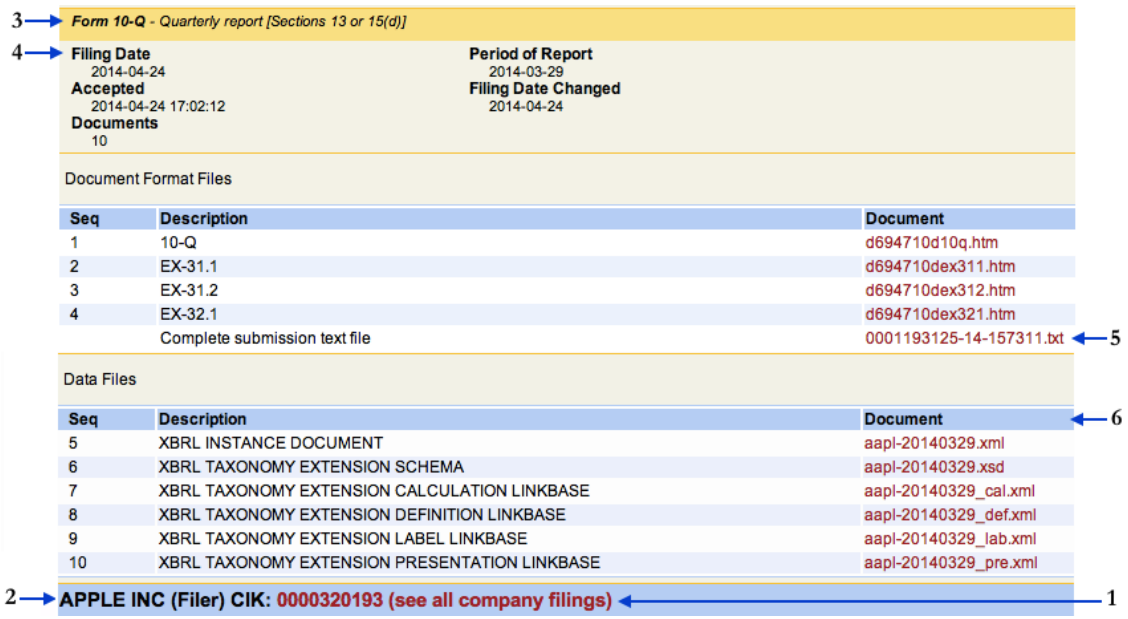
Todas las relaciones entre clases, subclases, taxonomías y valores, se han descrito de manera conceptual y forman parte de una base de conocimientos financieros fundamentada en Linked Data, cuya representación funcional requiere del procesamiento de ciertas aplicaciones de Software que permitan representar a cada una de las entidades que integran al modelo semántico en una base de conocimientos financieros inspirada en Linked Data realmente funcional. Con base en esto, en los párrafos siguientes se procede a describir el proceso de transformación a tripletas RDF de los datos almacenados en documentos XBRL publicados trimestralmente en el repositorio EDGAR y bajo la norma US-GAAP para la generación de la base de conocimientos financieros inspirada en los principios de Linked Data (T Berners-Lee, 2009) y cuyo núcleo es el modelo semántico de datos financieros que se ha venido describiendo en el transcurso del presente capítulo.

Primero, mediante un *Crawler* se rastrean y descargan los Estados financieros publicados por las empresas de manera trimestral, bajo el lenguaje XBRL y siguiendo la norma US-GAAP. El *Crawler* es ejecutado sobre el repositorio EDGAR y para su ejecución recibe dos parámetros, el primero consiste en un índice que contiene la información de los Estados financieros para el rastreo y descarga de los Estados financieros requeridos. El segundo parámetro, es la ruta del directorio en el que se almacenan las presentaciones de los Estados financieros XBRL *(XBRL filings)* descargados. Los índices son un ficheros con extensión .idx y se encuentran disponibles vía FTP (*File Transfer Protocol,* Protocolo de Transferencia de Archivos)[26] a través del sistema EDGAR.

Un ejemplo del formato de estos índices es el siguiente, `320193|APPLEINC|10-Q|2014-04-24|edgar/data/320193/0001193125-14-157311.txt` La Figura 17, muestra una sección de la interfaz Web correspondiente a este índice, y permite observar a detalle la

[26]Índices: ftp://ftp.sec.gov/edgar/

presentación de los Estados financieros publicados por la empresa APPLE INC. A través del sistema EDGAR de la U.S. SEC.



3 → ***Form 10-Q*** *- Quarterly report [Sections 13 or 15(d)]*

4 → **Filing Date** 2014-04-24 **Accepted** 2014-04-24 17:02:12 **Documents** 10 **Period of Report** 2014-03-29 **Filing Date Changed** 2014-04-24

Document Format Files

Seq	Description	Document
1	10-Q	d694710d10q.htm
2	EX-31.1	d694710dex311.htm
3	EX-31.2	d694710dex312.htm
4	EX-32.1	d694710dex321.htm
	Complete submission text file	0001193125-14-157311.txt ← 5

Data Files

Seq	Description	Document ← 6
5	XBRL INSTANCE DOCUMENT	aapl-20140329.xml
6	XBRL TAXONOMY EXTENSION SCHEMA	aapl-20140329.xsd
7	XBRL TAXONOMY EXTENSION CALCULATION LINKBASE	aapl-20140329_cal.xml
8	XBRL TAXONOMY EXTENSION DEFINITION LINKBASE	aapl-20140329_def.xml
9	XBRL TAXONOMY EXTENSION LABEL LINKBASE	aapl-20140329_lab.xml
10	XBRL TAXONOMY EXTENSION PRESENTATION LINKBASE	aapl-20140329_pre.xml

2 → **APPLE INC (Filer) CIK: 0000320193 (see all company filings)** ← 1

Figura 17. Ejemplo de presentación de los Estados financieros de APPLE INC. (U.S. SEC., 2013, 2014a)

Con base en la numeración asignada en la Figura 17, se describen brevemente los principales elementos que conforman la estructura del índice ejemplificado.

1. El CIK (*Central Index Key,* Clave de Índice Central) se utiliza en los sistemas informáticos de la U. S. SEC, para identificar a las empresas que han presentado la divulgación de sus Estados financieros ante la organización con el mismo nombre.
2. El nombre de la empresa que publica sus Estados financieros, para este ejemplo es APPLE INC.
3. 10-Q corresponde al formato trimestral de los Estados financieros publicados por parte de las empresas.
4. *Filing Date* es la fecha en la que las empresas presentan sus Estados financieros.
5. El fichero de presentaciones *(Complete submission text file)*, es un fichero .txt que contiene la información (la ruta) para la obtención de las taxonomías y la información financiera a través de FTP en la Web del sistema EDGAR.
6. Las empresas presentan sus Estados financieros en XBRL siguiendo una estructura de taxonomías y e información financiera basada en documentos XML y esquemas XSD. Esta estructura fue descrita en la sección 2.3.3 de la presente tesis doctoral.

Todos los documentos XBRL mostrados en la sección *Data files* (6) de la Figura 17 son descargados por el *Crawler*. Particularmente, el documento de mayor interés es el *XBRL*

INSTANCE DOCUMENT porque contiene los datos financieros de la empresa, es decir, los ratios, sus valores, e información adicional como el DEI (*Document Information and Entity Information,* Información del documento e información de la Entidad) cuya descripción es proporcionada más adelante. Dentro del modelo semántico, este documento está representado por *"Financial statement"* que es publicado por las empresas (En el ejemplo, APPLE INC). Hay que subrayar que el modelo semántico fomenta la reutilización de taxonomías y vocabularios con la finalidad de mantener los estándares ya establecidos, en este sentido, las taxonomías XBRL descargadas son reutilizadas para la generación de la base de conocimientos financieros incluyendo la reutilización de vocabularios como *Time Ontology* para especificar el instante *(instant)* y el periodo *(period)* (Fecha inicial y Fecha final) en el que fueron publicados los Estados financieros, y el vocabulario *Payments Ontology* con el propósito de especificar la unidad de medida aplicada a los valores numéricos, y que para cada estado financiero, está representada en la taxonomía *"Financial Unit Taxonomy"*. Siguiendo este contexto, la unidad de medida corresponde al tipo de moneda utilizada en los conceptos financieros y es publicada bajo el estándar *Currency Codes* - ISO 4217 (ISO, 2008a).

Los Estados financieros de Hoja de balance, Estado de flujo de efectivo y Cuenta de estado de resultados que integran al modelo semántico, son documentos transformados en tripletas RDF. En el modelo semántico, se muestra el ejemplo de una implementación de la taxonomía de Hoja de balance *("Balance sheet")* en la que se categorizan los ratios financieros en dos grupos principales, *"Assets"* y *"Liabilities and Equity"*. De la Figura 13 a la Figura 13.8, se muestran los ratios financieros contenidos en esta taxonomía. Es importante mencionar que los ratios que conforman a las taxonomías, en especial a la Hoja de balance fueron analizados por un grupo de expertos en finanzas.

Una vez almacenados los Estados financieros XBRL requeridos, y tras desarrollar las taxonomías financieras necesarias, se continua con la extracción y transformación hacia tripletas RDF de los datos contenidos en los Estados financieros XBRL descargados. Para realizar este proceso, fue necesaria la ejecución de un Extractor de datos en el que técnicamente se importan las taxonomías financieras que integran al modelo semántico para la generación del modelo de datos que da origen a la transformación en tripletas RDF de los datos financieros. Posteriormente, se inicia con la extracción y transformación de la información del DEI contenida en el documento *XBRL INSTANCE DOCUMENT*. Esta información también es presentada en formato XBRL y en ella que se combinan los datos

de la empresa solicitante de registro y la información de los documentos que esta pone de manifiesto. Para distinguir esta información, los documentos presentados por la empresa incluyen datos como el Tipo de documento *(Document Type)*, la Fecha final del periodo de publicación *(Period End Date)*, el año fiscal *(Fiscal Year Focus)* y el período fiscal *(Fiscal Period Focus)* de los Estados financieros publicados, entre otros datos correspondientes a estos documentos. Por otra parte, la información de la empresa solicitante incluye datos como el CIK, el Nombre de la entidad registrante *(Entity Registrant Name)* y el *Ticker Symbol*, por mencionar algunos datos de la empresa. Retomando el ejemplo de APPLE INC., en la Figura 18, se muestra una tabla con la información del DEI correspondiente a los Estados financieros publicados por esta empresa para el periodo del 29 de Marzo de 2014.

Document and Entity Information	6 Months Ended	
	Mar. 29, 2014	Apr. 11, 2014
Document Type	10-Q	
Amendment Flag	false	
Document Period End Date	Mar. 29, 2014	
Document Fiscal Year Focus	2014	
Document Fiscal Period Focus	Q2	
Trading Symbol	AAPL	
Entity Registrant Name	APPLE INC	
Entity Central Index Key	0000320193	
Current Fiscal Year End Date	--09-27	
Entity Filer Category	Large Accelerated Filer	
Entity Common Stock, Shares Outstanding		861,381,000

Figura 18. DEI de APPLE INC. Para el periodo del 29 Marzo de 2014 (U.S. SEC., 2013, 2014b)

Dentro del modelo semántico, todos los datos del DEI están contenidos en la clase "*Financial statement"* y son descritos en la taxonomía *"XBRL Statements Taxonomy"*. La transformación de estos datos en RDF, es incluida en los datos de la clase *"Balance sheet"* del modelo semántico. Después de extraer y transformar los datos del DEI de la empresa, el Extractor de datos obtiene y transforma en tripletas RDF los ratios y valores contenidos en clase "*Financial statement",* que como ya se mencionó es la representación conceptual del documento *XBRL INSTANCE DOCUMENT*. Los ratios extraídos y transformados a partir de este documento, también se incluyen en la clase *"Balance sheet"* y están representados por la clase *"Balance sheet field"* clasificándolos en *"Assets"* y *"Liabilities Equity"*. Con relación a los ratios financieros, la extracción y transformación de sus respectivos valores, es representada mediante la clase *"Value"*.

Es importante mencionar que el Extractor de datos, además de extraer y transformar en tripletas RDF los datos contenidos en documentos XBRL, también los transforma en notación *Turtle* con el objetivo generar una base de conocimientos financieros con un grafo alternativo basado en el modelo Entidad-Atributo-Valor e inspirado en los principios de Linked Data (T Berners-Lee, 2009). Este modelo es descrito en la sección siguiente.

4.5 Modelo Entidad-Atributo-Valor de datos financieros inspirado en Linked Data

Las ventajas y características que ofrece el modelo de datos Entidad-Atributo-Valor, también conocido como modelo EAV, son explotadas con el objetivo de proporcionar una base de conocimientos alternativa a la generada con el modelo Mixto de datos financieros descrito en la sección anterior, el cual posee características del modelo EAV y diseño canónico. A diferencia de este, el modelo EAV se complementa con metadatos añadidos a través de la reificación cuya descripción se encuentra al final de la sección 2.4.4 de la presente tesis.

La Figura 19, muestra la conceptualización del modelo EAV de datos financieros con reificación.

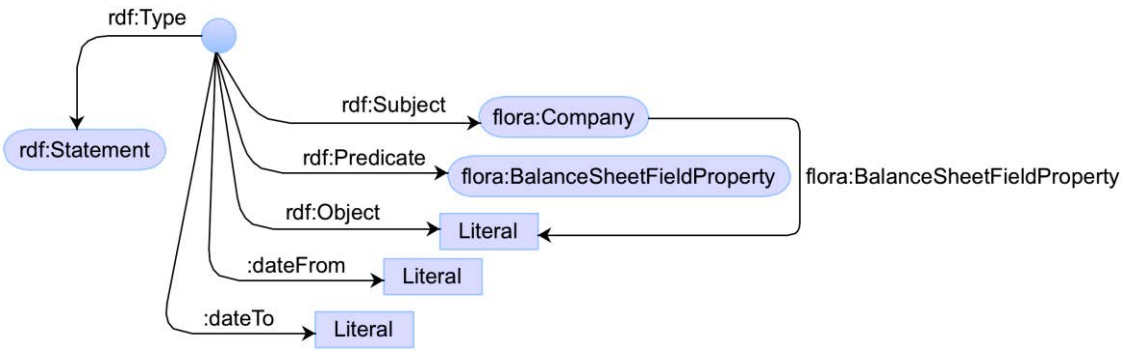


Figura 19. Conceptualización del modelo EAV de datos financieros con reificación

La conceptualización del modelo EAV es más simple en comparación con el modelo semántico Mixto. Mientras que este último describe los datos de manera normalizada y en forma de clases, taxonomías y valores relacionados, e implementa los valores que definen el periodo de publicación de los Estados financieros XBRL en las subclases *"Period"* e *"Instant"*, el modelo EAV a través de la reificación, representa estos valores como metadatos añadidos (nodos blancos) a las tripletas RDF generadas a partir de la extracción y transformación de los Estados financieros XBRL. En este sentido, el Extractor de datos genera las tripletas RDF en sintaxis *Turtle* bajo la siguiente estructura:

- **Entidad *(Entity)*:** es el nombre de la empresa.
- **Atributo *(Attribute)*:** estos corresponden a los ratios financieros contenidos en el documento *XBRL INSTANCE DOCUMENT* y sus taxonomías financieras (Hoja de balance, Cuenta de estado de resultados y Estado de flujo de efectivo).
- **Valor *(Value)*:** es el valor asignado al ratio financiero representado por los Atributos.
- **Nodo blanco *(Blank Node)*:** se generan dos nodos blancos que indican la fecha de inicio y fin del periodo de publicación del estado financiero.

La conceptualización de los dos modelos de datos semánticos presentados en este capítulo son útiles para la generación de la base de conocimientos financieros inspirada en Linked Data. Sin embargo, ambos presentan notables diferencias tanto en su representación de datos como en su navegabilidad, tocado este punto, en la sección siguiente se continua con la descripción de la navegabilidad de cada uno de estos modelos.

4.6 Navegabilidad de los modelos semánticos

La adquisición de datos financieros está directamente relacionada con el propósito para el que han sido diseñados los modelos semánticos inspirados en los principios de Linked Data descritos a lo largo del presente capítulo. Siguiendo este orden de ideas, el modelo EAV presenta características para la navegación más directa entre los datos y para la búsqueda de la información, ya que esta se realiza mediante un sólo paso, tal y como se muestra en la Figura 20. Mientras que el modelo semántico Mixto tiene características de normalización que favorecen la realización de cálculos matemáticos y la búsqueda de información, pero no de una manera tan directa, como se muestra en la Figura 21.

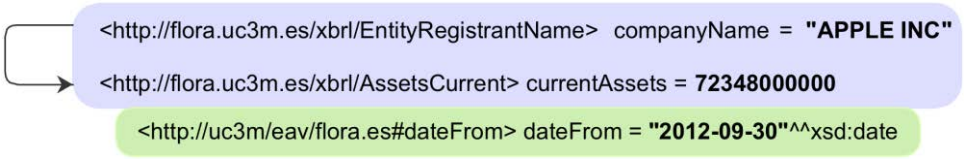


Figura 20. Representación de navegabilidad en el modelo EAV

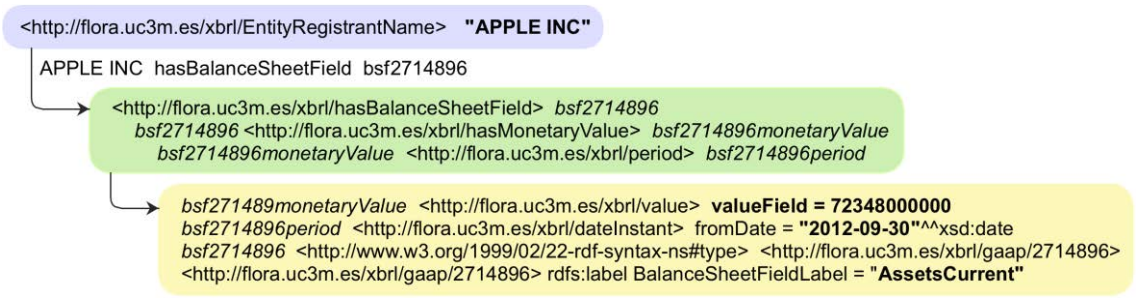


Figura 21. Representación de navegabilidad en el modelo semántico Mixto

Para ambos modelos, las Figuras 20 y 21 representan un ejemplo de navegabilidad para buscar el valor asignado al ratio financiero *"Current Assets"* en una fecha de terminada para la empresa "APPLE INC". En el caso del modelo EAV, se muestra que la búsqueda de esta información se recupera en un sólo paso, en el que el nombre de la empresa es la Entidad, el *"AssetsCurrent"* corresponde al Atributo y 72348000000, es el Valor asignado al Atributo. Además, este modelo incluye la integración del *"dateFrom"* como un metadato añadido a través de la reificación.

A diferencia del modelo EAV, el modelo Mixto realiza la búsqueda y recuperación de datos mediante tres pasos. En el primer paso se recupera el nombre de la empresa, en el siguiente paso la empresa se relaciona con su *"moneataryValue"* y su *"period",* a través de un bsf-id *(balanceSheetField-id),* cuyo valor ejemplificado en la Figura 21 es bsf2714896. Finalmente, los valores son recuperados en el tercer paso, el *"monetaryValue"* obtiene su valor correspondiente y *"period"* obtiene el valor *"fromDate"*, basándose en el criterio de búsqueda *"AssetsCurrent"*, que se recupera a través de *"BalanceSheetFieldLabel"*.

Desde una perspectiva de representación normalizada de datos, el modelo Mixto presenta las clases, las relaciones y los valores que permiten una fácil interconexión e integración entre estos, lo que proporciona un procesamiento más orientado hacia el cálculo de los datos y por ende, hacia su fácil interpretación por parte de la máquina. A diferencia del modelo Mixto, el modelo EAV es más transparente, ya que mantiene un esquema abierto hacia la integración de datos centrados en la Entidad. Esto significa, que la obtención de los atributos y sus valores se realiza a través de la Entidad (Nombre de la empresa). Esta característica es fundamental para el tratamiento de datos en la Web, porque favorece a la navegación y la búsqueda de información (Freitas et al., 2013). Por lo tanto, el modelo EAV es un modelo con orientación hacia la fácil interpretación humana.

Los modelos semánticos de datos financieros inspirados en Linked Data presentados a lo largo de este capítulo, mantienen distintas características. Sin embargo, ambos modelos

son alternativas viables que permiten la generación de una base de conocimientos financieros que sirve de apoyo para el cumplimiento de los objetivos descritos en el presente trabajo de tesis y que principalmente, da soporte a la validación de las hipótesis planteadas. Elegir el mejor modelo semántico para la publicación estructurada de datos financieros no es una tarea trivial. Por lo tanto, antes de iniciar con los experimentos para elegir y validar el modelo semántico que mejor se adapte a los objetivos e hipótesis planteados en esta tesis, en el Capítulo 5 se describen las estadísticas de la base de conocimientos financieros inspirada en Linked Data, con la finalidad de conocer más a detalle el número de elementos que la integran.

Capítulo 5

Base de conocimientos financieros inspirada en Linked Data

Resumen. El capítulo anterior, consistió en proporcionar información acerca de los modelos semánticos inspirados en Linked Data para la publicación de datos financieros, sus taxonomías, la transición de su conceptualización hacia una base de conocimientos y su navegabilidad. En este sentido, cada uno de los elementos que integran los modelos semánticos previamente descritos, son indispensables para generar una base de conocimientos que técnicamente sirva de apoyo para la el procesamiento de los experimentos que permitan validar las hipótesis planteadas en el Capítulo 3 de este trabajo de tesis. Sin embargo, antes de dar inicio con los experimentos, en el presente capítulo se proporciona información de las estadísticas concernientes a la base de conocimientos financieros generada, con el fin de cuantificar los Estados financieros que han sido descargados y transformados en tripletas RDF, así como conocer el tipo de datos con los que se realizará la experimentación.

5. Base de conocimientos financieros inspirada en Linked Data

La estadística como ciencia representa un extenso campo de estudio, sin embargo, limitándose a la estadística descriptiva (Trochim, 2006), el significado del término "estadísticas" se reduce a la selección de datos numéricos presentados en forma esquemática y ordenada. En este sentido, las estadísticas favorecen la elaboración de encuestas, organización datos, tabulación, representaciones y cálculo de parámetros (Mann, 1995; Trochim, 2006).

Sustentándose en lo descrito en el párrafo anterior, las estadísticas de la base de conocimientos financieros tienen un papel importante para los fines de validación que se presentan en el Capítulo 6 del presente trabajo de tesis, por ejemplo, apoyan con la cuantificación del número total de los Estados financieros XBRL descargados, el número total de tripletas RDF transformadas y almacenadas en la base de conocimientos, sus clases, sujetos, propiedades y objetos. Dicho lo anterior, en las secciones siguientes de este capítulo se proporciona información acerca de los Estados financieros XBRL descargados, la infraestructura tecnológica utilizada, estadísticas de los ficheros RDF generados tras la transformación de los Estados financieros XBRL y las estadísticas de las tripletas RDF que integran a la base de conocimientos.

5.1 Estadísticas de los Estados financieros XBRL descargados para su transformación en RDF

Retomando la fase de Adquisición de datos descrita en el proceso para la generación de la base de conocimientos financieros inspirada en Linked Data (véase Sección 3.2), se desarrolló un *Crawler*, también conocido como Araña Web con tecnología Java que permite rastrear e identificar los Estados financieros XBRL publicados por las empresas de manera trimestral, es decir, en formato 10-Q y que siguen la norma US-GAAP. La fuente principal de estos Estados financieros es el repositorio EDGAR de la U.S. SEC. Los Estados financieros son descargados y almacenados para la posterior extracción y transformación de sus datos en tripletas RDF tomando como modelo de datos, los modelos semánticos Mixto y EAV previamente descritos en el Capítulo 4 de este trabajo de tesis.

La Tabla 6, presenta la información concerniente a los conjuntos de Estados financieros XBRL *(filings),* descargados a través del *Crawler* a partir del año 2009 hasta el tercer trimestre de 2014.

ESTADOS FINANCIEROS RASTREADOS Y DESCARGADOS					
AÑO	TRIMESTRE				TOTAL TRIMESTRES
	Q1	Q2	Q3	Q4	
2009	NA	504	9,223	13,831	23,558
2010	2,255	14,162	45,012	51,909	113,338
2011	10,983	56,976	38,813	47,548	154,320
2012	13,259	48,767	41,077	46,669	149,772
2013	12,968	46,927	343,097	46,164	449,156
2014	12,375	42,777	49,400	NA	104,552
			TOTAL INFORMES DESCARGADOS:		**994,696**
Abreviaturas: NA.- No Aplica; Q.- Número de Trimestre.					

Tabla 6. Detalle de los Estados financieros descargados hasta el tercer trimestre de 2014

Entre los documentos XBRL descargados, se encuentran los Estados financieros de Hoja de balance, Estado de flujo de efectivo y Cuenta de estado de resultados, así como esquemas XML y ficheros adicionales que integran la estructura de los Estados financieros. Como ya se ha descrito, los documentos XBRL descargados con el *Crawler* son transformados en tripletas RDF, lo que da origen a un determinado número de tripletas de este tipo que son almacenadas en el repositorio Virtuoso Open-Source. De acuerdo con lo expresado, en la sección siguiente se proporciona información general acerca de la infraestructura tecnológica utilizada para este repositorio semántico.

5.2 Infraestuctura tecnológica utilizada para el almacenamiento de las tripletas RDF transformadas

Mediante un extractor de datos desarrollado en Java y Apache Jena, los datos contenidos en los Estados financieros XBRL son extraídos y transformados en tripletas RDF según el modelo de datos que adopte. Esto significa que primero se realiza una transformación en tripletas RDF con base en el modelo semántico Mixto y posteriormente, se realiza una transformación en tripletas del mismo tipo, pero adaptadas al modelo EAV con sintaxis *Turtle* y metadatos añadidos a través de reificación.

La finalidad de realizar estas transformaciones, es la de integrar una base de conocimientos financieros con dos grafos distintos. En otras palabras, cada modelo semántico de datos financieros tiene asignado un grafo diferente mediante el que se accede directamente a sus datos a través de consultas basadas en SPARQL. Como ya se ha descrito previamente en la fase de Generación del conjunto de datos semántico (véase Sección 3.2), las tripletas RDF se almacenan en el repositorio Virtuoso Open-Source

(Erling & Mikhailov, 2009). Siguiendo este contexto, en la Tabla 7 se muestra información general de la infraestructura tecnológica utilizada para este repositorio y el número de ficheros RDF transformados para cada modelo semántico.

INFRAESTRUCTURA TECNOLÓGICA			
Sistema Operativo	**RAM**	**Procesador**	**Tipo de sistema**
Windows Server 2008 R2 Standard Service Pack 1	RAM 8 GB	Processor AMD Phenom(tm) II X6 1090 3.20GHz	64 bits
PLATAFORMA SPARQL ENDPOINT			
Repositorio semántico		**Versión - Build date**	
Virtuoso Open-Source Edition		06.01.3127 - Agosto 2 - 2012	
Dirección URL			
http://nadir.uc3m.es:8890/sparql			
Grafo modelo semántico Mixto		**Grafo modelo semántico EAV**	
http://flora/mixed/uc3m.es		http://flora/eav/uc3m.es	
TIEMPO DE CARGA DE LAS TRIPLETAS RDF EN EL REPOSITORIO SEMÁNTICO			
Modelo semántico Mixto		**Modelo semántico EAV**	
73, 727 Ficheros (11,1 GB)	7h 9min.	73, 724 Ficheros (16,2 GB)	9h 18min.

Tabla 7. Infraestructura tecnológica utilizada para el repositorio de datos financieros

La decisión de utilizar el repositorio Virtuoso Open-Source, se sustenta en que es la plataforma para la gestión, acceso, consulta e integración de datos basados en el paradigma Linked Data más conveniente para llevar a cabo los experimentos que sirven de apoyo para la validación de las hipótesis planteadas en el presente trabajo de tesis. Esta decisión se fundamenta en los resultados obtenidos por Bizer & Schultz, (2009), quienes compararon varios sistemas para la gestión de datos basados en Linked Data mediante pruebas de referencia *(benchmark)*, en las que procesaron la carga de datos RDF y ejecutaron un conjunto de consultas basadas en SPARQL, teniendo como conclusión, que con Virtuoso Open-Source obtenían los mejores resultados.

Para corroborar esta decisión, Morsey et al., (2011), realizó pruebas de referencia sobre DBPedia través de consultas basadas en SPARQL procesadas en Virtuoso Open-Source, Sesame, Apache Jena-TDB y BigOWLIM. Los resultados obtenidos por el autor y sus colaboradores indican claramente que Virtuoso Open-Source fue el más rápido. Finalmente, los autores demostraron que no todas las consultas basadas en SPARQL podrían ser procesadas dentro de un tiempo de espera de 180 segundos, y se dieron cuenta que aplicar tiempos de espera aún mucho más grandes no habrían sido suficientes para procesar la mayoría de esas consultas. Con excepción de Virtuoso Open-Source, que fue la

única plataforma capaz de procesar todas las consultas basadas en SPARQL dentro de los 180 segundos.

Las estadísticas originadas por la transformación en tripletas RDF de los datos contenidos en los Estados financieros XBRL ayudan a conocer más a detalle los datos que se encuentran almacenados en la base de conocimientos financieros inspirada en Linked Data. Por este motivo, en la sección siguiente se obtienen las estadísticas de las tripletas RDF que integran a dicha base de conocimientos.

5.3 Estadísticas de las tripletas RDF almacenadas en la base de conocimientos financieros

El repositorio Virtuoso Open-Source almacena la información de los Estados financieros de las empresas en forma de tripletas RDF, esta información puede ser obtenida mediante consultas basadas en SPARQL, que se ha propuesto como el lenguaje de consultas estándar para la Web Semántica por ser intuitivamente comprensible (Dimitrov, 2012; Kobayashi & Toyoda, 2008). De acuerdo con lo dicho, se procesó una serie de consultas basadas en SPARQL que calculan algunas estadísticas concernientes a la base de conocimientos financieros. La primer consulta consiste en obtener el número total de tripletas RDF almacenadas en la base de conocimientos financieros, los resultados obtenidos se muestran en la Figura 22 en forma de porcentajes para cada modelo semántico.

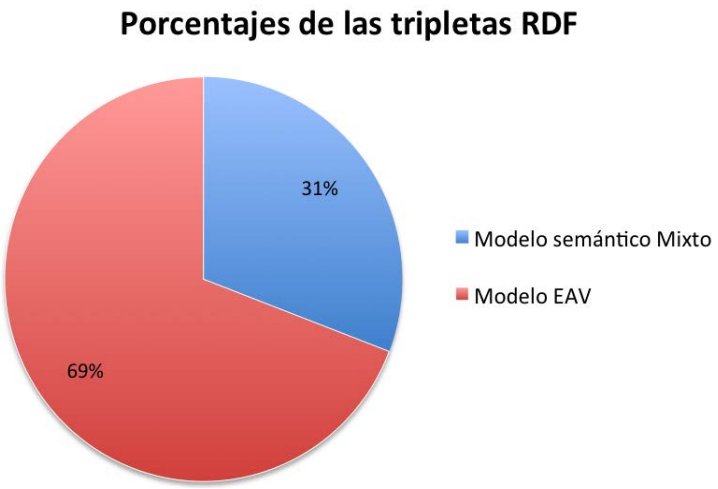


Figura 22. Porcentajes correspondientes a las tripletas RDF de la base de conocimientos financieros

Hasta el tercer trimestre de 2014, el 100% de la base de conocimientos financieros inspirada en Linked Data contiene un total de 343,244,089 tripletas RDF de las que, 105,787,711 corresponden al modelo semántico Mixto, lo que equivale al 31% del total de

las tripletas. Por otra parte, 237,456,378 tripletas RDF corresponden al modelo EAV y conforman el 69% del total de las tripletas. Ambos modelos semánticos representan la misma información extraída de los Estados financieros XBRL descargados con el *Crawler*. Sin embargo, la transformación en tripletas RDF tiene un formato distinto para cada modelo semántico, por ejemplo, las tripletas RDF transformadas con base en el modelo EAV contienen metadatos añadidos mediante reificación para incluir el periodo de publicación de los Estados financieros, por este motivo, el número de tripletas es mayor en comparación con el modelo semántico Mixto que incluye la misma información de los periodos de publicación de los Estados financieros pero a través de la clase *Period*.

Seguido de la obtención del número total de tripletas RDF almacenadas en la base de conocimientos financieros, se procesaron consultas adicionales basadas en SPARQL para obtener estadísticas sobre las tripletas RDF que conforman a cada modelo semántico. Los resultados indican que las tripletas RDF del modelo semántico Mixto se componen de 1,516 clases, 29,608,497 nodos sujeto, 133 predicados y 29,388,808 nodos objeto distintos. Entre las clases que conforman las tripletas RDF de este modelo semántico se encuentran *Company, Period, BalanceSheet, BalanceSheetField, StatementOfIncome, StatementOfIncomeField, StatementOfCashFlows, StatementOfCashFlowsField* y *MonetaryValue,* por mencionar las de mayor importancia.

Contrario a las estadísticas obtenidas mediante la ejecución de las consultas basadas en SPARQL sobre las tripletas RDF que conforman el modelo semántico Mixto, los resultados de las estadísticas de las tripletas RDF del modelo EAV muestran que estas tripletas se componen por 1 tipo de clase, 36,255,918 nodos sujeto, 10,864 predicados y 22,898 nodos objeto distintos. A diferencia del modelo semántico Mixto que se compone de varias clases, el modelo EAV maneja clases de tipo rdf:Statement que es una instancia de rdfs:Class (Brickley & Guha, 2014).

Como información complementaria a las estadísticas previamente descritas, el número total de tripletas RDF almacenadas en la base de conocimientos financieros se compone de 1,517 clases, 65,864,415 nodos sujeto, 10,997 predicados y 29,411,706 nodos objeto distintos. Finalmente, la información proporcionada en el presente capítulo, es importante para tener una estimación cuantificada del número de elementos que integran a la base de conocimientos financieros inspirada en Linked Data generada a partir de los modelos semánticos que se han descrito a lo largo de este trabajo de tesis doctoral. Pero principalmente, para conocer los tipo de datos con los que se realizarán los experimentos

que permitan validar las hipótesis planteadas en el Capítulo 3 de este documento. Dicho esto, en el capítulo siguiente se describe a detalle cada uno de los experimentos realizados para la validación de las hipótesis, así como la descripción de los resultados obtenidos.

Capítulo 6

Validación de los modelos semánticos de datos financieros inspirados en Linked Data

Resumen. Una vez descritos los modelos Mixto y EAV inspirados en Linked Data, así como las estadísticas e información general de la base de conocimientos generada a partir de estos modelos semánticos, en el presente capítulo se da inicio con los experimentos de validación necesarios que permitan identificar, elegir y justificar el modelo semántico que mejor se ajusta a los propósitos que con esta tesis doctoral se persiguen. Con base en los resultados obtenidos, se procede con en el diseño y ejecución de diversos experimentos aplicados sobre el modelo semántico elegido con el objetivo de validar las hipótesis planteadas en el Capítulo 3 del presente trabajo de tesis.

6. Introducción y validación de los modelos semánticos de datos financieros inspirados en Linked Data

Para que un proceso de investigación sea considerado como de muy importante, es necesario que este proceso se encuentre estrechamente relacionado con el mundo real de aplicación de su sector (Eco, 2001). En este orden de ideas, el presente trabajo de tesis se orienta hacia el sector financiero para el que, como ya se ha descrito en el apartado 1.2 de Justificación, se propone un modelo semántico inspirado en los principios de Linked Data (T Berners-Lee, 2009) que sirva de alternativa de solución a los problemas de integración de datos que presentan los Estados financieros basados en el estándar XBRL publicados por las empresas.

Durante el transcurso de una investigación existe la posibilidad de que esta avance por una trayecto que diverja de la realidad práctica y científica que le dio origen, dificultando la aplicación de los resultados y conclusiones obtenidas. Por ello, es necesario conocer y considerar las inquietudes, opiniones y tendencias de los profesionales del sector sobre el que se realiza el trabajo de investigación (González & Padilla, 1999).

Para el presente trabajo de tesis, se optó por diseñar y ejecutar un conjunto de experimentos, cuyos resultados faciliten la validación de las hipótesis planteadas y el cumplimiento de los objetivos estipulados. Los experimentos a diseñar y ejecutar, tienen la intención de proporcionar resultados que validen cualitativa o cuantitativamente cada una de las hipótesis planteadas. Entre las validaciones consideradas, se encuentra la participación de expertos en finanzas y contabilidad a través de un cuestionario en el que sus respuestas y opiniones ayudarán a constatar que los resultados obtenidos pueden ser, o no, de utilidad para el sector al que está orientado este trabajo de tesis.

Con la validación de los modelos semánticos inspirados en Linked Data para la publicación de datos financieros en la Web, se persiguen algunos objetivos que ayudarán a sustentar los resultados y conclusiones obtenidas. Estos objetivos son los siguientes:

- Diseñar y ejecutar los experimentos necesarios para validar cada una de las hipótesis planteadas en este trabajo de tesis.
- Constatar que los resultados obtenidos para cada experimento, permiten la correcta validación de las hipótesis planteadas.
- Comprobar la propuesta de validación de las hipótesis descrita en la Sección 3.1

- Confirmar con expertos en el sector financiero los beneficios que se pueden obtener con la información contenida en la base de conocimientos inspirada en Linked Data.
- Verificar el cumplimiento de los objetivos general y específicos descritos en el capítulo inicial del presente documento.

Determinar los objetivos a seguir con las validaciones a realizar, proporciona la pauta para dar inicio con el diseño y ejecución de los experimentos que faciliten el cumplimiento de los mismos. Dicho esto, los experimentos a realizar se describen en las secciones siguientes iniciando con la realización de un análisis comparativo entre los modelos semánticos Mixto y EAV.

6.1 Análisis comparativo entre los modelos semánticos Mixto y Entidad-Atributo-Valor

El modelado semántico desempeña un papel central en todo sistema basado en conocimiento, donde el intercambio de información y la integración de datos es un objetivo primordial. La adopción de los grafos RDF como estructura para la representación semántica de metadatos permite la ejecución de consultas sencillas y expresivas utilizando el lenguaje SPARQL.

Mientras que el rendimiento de los sistemas basados en conocimiento depende de múltiples factores como las características de Hardware que los albergan, en esta validación se describe que la elección adecuada del patrón de modelado semántico puede reducir significativamente los tiempos de procesamiento en las consultas de datos. Sobre la base de esta comprensión, en la presente sección se describe un análisis comparativo *(benchmarking)* entre los modelos semánticos presentados durante esta investigación, teniendo como base el dominio financiero y con la intención de elegir el modelo que mejor convenga a los propósitos que con este trabajo de tesis se persiguen.

6.1.1 Diseño de consultas basadas en SPARQL para el análisis comparativo entre los modelos semánticos

Con el fin de mostrar la eficiencia y precisión de este enfoque de validación, se diseñó un conjunto de consultas basadas en SPARQL contra la base de conocimientos generada a partir de los modelos semánticos Mixto y EAV. El objetivo de la ejecución de estas consultas (véase Tabla 8), es conocer los tiempos de adquisición de datos para cada modelo

semántico y de esta manera identificar y elegir el modelo que mejor se adecúa para la búsqueda de datos y la realización de operaciones de tipo financiero, pero que principalmente, permita el cumplimiento de los objetivos y la validación de las hipótesis planteadas en este trabajo de tesis.

CONSULTAS SPARQL	DESCRIPCIÓN
Q1	Recuperar toda la información de los primeros 500,000 registros almacenados en el conjunto de datos financieros.
Q2	Obtener una lista con el nombre de las empresas y sus correspondientes CIK's *(Central Index Key)*, registrados en el conjunto de datos financieros.
Q3	Obtener los ratios financieros pertenecientes a la empresa "Apple Inc.", indicando el valor, el trimestre y la fecha de publicación para cada concepto financiero.
Q4	A partir de las empresas "GOOGLE INC, MICROSOFT CORP Y YAHOO INC", recuperar la información de los ratios de sus Hojas de balance con sus respectivos valores, cuya fecha de publicación está entre el 01/01/2011 y el 31/12/2014.
Q5	Obtener la información de las empresas cuyo valor de plusvalía o valor de buena voluntad *(Goodwill Value)* es mayor que 1000.000.000 dls. Para los Estados financieros publicados sólo en el año 2013.
Q6	Obtener el valor promedio de los Activos Corrientes *(Current Assets)* para la empresa "COCA COLA CO", para los Estados financieros publicados a partir del 01/01/2014
Q7	Consultar el valor mínimo de los Pasivos Corrientes *(Current Liabilities)* registrados para la empresa "GENERAL MOTORS COMPANY", entre las fechas del 01/01/2013 y 31/07/2014.
Q8	Calcular el valor de Prueba ácida *(Acid Test)* para la empresa "WAL MART STORES INC", basándose en el año fiscal en el que fueron publicados sus Estados financieros. La Prueba ácida, es una relación de contabilidad que indica la liquidez o solvencia de una empresa en el corto plazo (Montero & Fernández-Aviles, 2010). Fórmula: *Acid Test = (Current Assets - Inventory) / Current Liabilities*
Q9	Calcular el valor del Capital de trabajo Neto Sobre Deudas a Corto Plazo para la empresa "PFIZER INC". Cuando el resultado es cercano 0,5 se considera como un nivel óptimo, pero si el valor es menor que 0,5 es posible que la empresa tenga problemas para cumplir con sus deudas a corto plazo, aunque convierta en dinero todos sus activos (Montero & Fernández-Aviles, 2010). Fórmula: *Net working Capital on Short-Term Debt = (Current Assets - Current Liabilities) / Current Liabilities*
Q10	Calcular los Días de Medición del Intervalo de Tiempo *(Day Time Interval Measurement)* para la empresa "ABTECH HOLDINGS, INC", con base en el año fiscal 2012. Este indicador financiero, permite obtener el número de días en los que una empresa puede seguir funcionando, si por alguna razón, se paralizan sus actividades diarias (Montero & Fernández-Aviles, 2010). Fórmula: *Day time interval measurement = (Current Assets / Cost of Materials) * 365*

Tabla 8. Consultas basadas en SPARQL para el *Benchmarking* de los modelos semánticos

Las consultas basadas en SPARQL para la realización del análisis comparativo entre ambos modelos semánticos fueron diseñadas con una complejidad inicialmente baja, en las que se aplican filtros simples para la adquisición de datos, posteriormente, la complejidad se incrementa a través del uso de filtros en los que se incluyen rangos de fechas, valores y funciones de agregado. Finalmente, las últimas tres consultas comprenden el cálculo de indicadores financieros adicionales, siendo estas las de mayor complejidad dentro del conjunto de consultas.

Cada una de las consultas diseñadas, tiene características que favorecen la adquisición de los datos requeridos y facilitan el análisis de los resultados obtenidos después de su ejecución. En la Tabla 9, se presentan las características utilizadas en las consultas basadas en SPARQL procesadas para cada modelo semántico.

CARACTERÍSTICAS	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Forma de consulta SELECT	√	√	√	√	√	√	√	√	√	√
Modificador LIMIT	√									
Modificador ORDER BY				√	√					
Modificador DISTINCT		√	√	√	√			√	√	√
Función de agregado AVG						√				
Función de agregado Min							√			
Tipos de datos XMLSchema (xsd)				√	√	√	√	√	√	√
Operadores relacionales			√	√	√	√	√	√	√	√
Operadores lógicos				√	√		√	√		
Operadores aritméticos								√	√	√

Tabla 9. Características de las consultas basadas en SPARQL utilizadas en el *benchmarking*

Las consultas basadas en SPARQL diseñadas para el análisis comparativo entre los modelos EAV y Mixto, son de la forma SELECT porque devuelven todas, o un subconjunto de variables coincidentes con el patrón de la consulta procesada. Enfatizar en esta descripción es importante porque además de la forma SELECT, el lenguaje SPARQL ofrece tres formas adicionales para consultar los datos almacenados en un repositorio de tripletas RDF. La forma CONSTRUCT, retorna un grafo RDF construido mediante la sustitución de variables en un conjunto de plantillas de tripletas de este tipo, la forma ASK devuelve un valor booleano que indica si un patrón de consulta tiene o no solución, el resultado no devolverá información sobre las posibles soluciones de la consulta, porque sólo existe, o no una solución. Finalmente, la forma DESCRIBE devuelve un grafo RDF que proporciona una descripción acerca de los recursos encontrados (W3C, 2013).

Los experimentos realizados permiten analizar cómo las estrategias de modelado semántico afectan el rendimiento de la adquisición de datos a través de la ejecución de consultas basadas en SPARQL, esto sirve de ayuda para elegir el modelo más conveniente para validar las hipótesis planteadas y cumplir con los objetivos puntualizados en este trabajo de tesis. Por lo tanto, los resultados obtenidos son descritos en la sección siguiente.

6.1.2 Resultados del análisis comparativo entre los modelos semánticos

La métrica principal que se utilizó para comparar los resultados obtenidos en los modelos semánticos EAV y Mixto, es el tiempo de procesamiento para la adquisición de datos medido en milisegundos (ms). Para ambos modelos, se busca la misma información

mediante la ejecución de las consultas basadas en SPARQL descritas en la Tabla 8. Además, las consultas son ejecutadas cinco veces con el propósito de calcular el tiempo de procesamiento para la adquisición de los datos requeridos en cada una de ellas. El motivo por el que las consultas son ejecutadas de esta manera, radica en las posibles variaciones o alteraciones que pudieran existir durante su tiempo de procesamiento. El origen de estas variaciones se debe a la posibilidad de que el microprocesador del servidor se encuentre procesando otras tareas en el mismo instante en el que se ejecutan las consultas diseñadas para el análisis comparativo de los modelos semánticos.

Las consultas de los experimentos del análisis comparativo se ejecutan a través de la herramienta iSQL del repositorio Virtuoso Open-Source, los resultados obtenidos se descargan directamente en ficheros .txt, de esta manera, se evita el tiempo que conlleva el despliegue de los resultados en la interfaz de consola de la herramienta iSQL, o el tiempo de renderizado para mostrar los resultados en la interfaz Web del SPARQL endpoint.

Los resultados obtenidos después de la ejecución de las consultas se muestran en la Tabla 10, cuya descripción es proporcionada a continuación.

TIEMPO DE ADQUISICIÓN DE DATOS (Milisegundos/ms)										
MODELO	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
EAV	51995	2621	15288	6162	5755	811	811	10624	3650	750
	30607	1202	8081	2543	5753	265	390	9703	3550	570
	48126	1139	14680	2824	5750	421	328	9719	3550	750
	40639	1139	13899	1248	5753	421	359	10109	3657	550
	48469	1107	14430	2278	5750	406	203	9890	3550	731
Mixto	890205	4384	609	702	2277	421	375	5756	3370	718
	565581	1170	125	281	2418	218	234	2932	3400	530
	271676	2060	234	296	2262	202	218	3603	3104	702
	297291	1716	234	171	2387	202	234	3681	3370	530
	275826	1763	249	343	2309	234	188	3681	3510	749
TIEMPOS PROMEDIO DE ADQUISICIÓN DE DATOS (Milisegundos/ms)										
EAV	43967,2	1441,6	13275,6	3011,0	5752,2	464,8	418,2	10009,0	3591,4	670,2
Mixto	460115,8	2218,6	290,2	358,6	2330,6	255,4	249,8	3930,6	3350,8	645,8

Tabla 10. Tiempos de adquisición de datos para los modelo semánticos Mixto y EAV

Los primeros experimentos para el análisis comparativo se realizaron contra el modelo EAV. En la Tabla 10, se muestra que en este modelo, los tiempos promedio de adquisición de datos favorecen a las consultas Q2, Q4, Q6, Q7, Q9 y Q10 con valores menores a 5000ms. Mientras que el resto de las consultas, presentan tiempos promedio de entre 5752,2ms y 43967,2ms.

Si se clasifican las consultas por su complejidad y se analizan sus características (véase Tabla 9), se observa que en Q1 el tiempo promedio para la adquisición de datos es razonable, debido a que la cantidad de estos es muy grande, sin embargo, la única diferencia entre Q2 y Q3 estriba en que esta última incluye entre sus características "Operadores relacionales", lo que posiblemente ocasiona que su tiempo de adquisición de datos sea mayor. Por otra parte, en las consultas cuya complejidad es media (de Q4 a Q7), Q4 y Q5 tienen características análogas y sus tiempos promedio para la adquisición de datos son muy superiores a Q6 y Q7, que incluyen entre sus características "Funciones de agregado" para el cálculo del valor promedio y la búsqueda del valor mínimo de un determinado ratio financiero. Para finalizar el análisis de los tiempos promedio para la adquisición de datos en el modelo EAV, las consultas de complejidad alta Q8, Q9 y Q10 presentan características similares, sin embargo, el tiempo promedio para la adquisición de datos de Q8 es mayor en comparación con las dos últimas, siendo la inclusión de los "Operadores lógicos", la diferencia más notable entre las características de estas tres consultas.

Posterior al análisis de los resultados obtenidos en el modelo EAV, se continúa con la descripción de los resultados obtenidos en el modelo semántico Mixto. A diferencia del modelo EAV, los resultados obtenidos en el modelo Mixto demuestran que los tiempos promedio para la adquisición de datos de Q2 a Q10 son inferiores a 5000ms, lo que significa que son los tiempos más óptimos obtenidos después de la ejecución del conjunto de consultas basadas en SPARQL diseñadas para el análisis comparativo que se presenta en la presente sección. Sin embargo, si se comparan los tiempos promedio para la adquisición de datos en Q1, se observa que en esta consulta, el modelo EAV es el que mejores resultados obtiene. Los resultados alcanzados con la realización del análisis comparativo de los modelos semánticos EAV y Mixto, permiten justificar las conclusiones descritas en la sección siguiente.

6.1.3 Conclusión del análisis comparativo entre los modelos semánticos

La representación del conocimiento es la base para el intercambio y la reutilización de la información. De esta manera, las tecnologías semánticas y el paradigma Linked Data proporcionan las herramientas para la representación estructurada y el intercambio de conocimientos que permiten la recuperación de información basada en vocabularios comunes. Estas características son especialmente relevantes para el ámbito financiero, donde las fuentes de datos son diversas y es necesaria la existencia de un modelo apropiado para la representación y recuperación de información financiera. Basándose en esta

comprensión, las búsquedas, la navegación entre datos y el cálculo de ratios en el dominio financiero son particularmente relevantes, sin embargo, es importante considerar los problemas de rendimiento en la adquisición de datos, a fin de proporcionar la información correcta en el tiempo adecuado. Por este motivo, en la presente sección se diseñaron y ejecutaron una serie de consultas basadas en SPARQL para los modelos semánticos inspirados en Linked Data descritos a lo largo del presente trabajo de tesis.

Tanto el modelo EAV como el modelo Mixto, permiten la realización de búsquedas, la navegación entre datos y el cálculo de indicadores financieros, sin embargo, basándose en la descripción de los resultados presentados en la Tabla 10, y excluyendo a Q1, la suma del total de los tiempos promedio para cada modelo es de 38634,0ms y 13630,4ms respectivamente, teniendo una diferencia de 25003,6ms lo que beneficia al modelo semántico Mixto. Aunado a esto, y como se muestra en la Tabla 7, el tiempo de carga de las tripletas RDF para este modelo, es de 2h 9min menor en comparación con el modelo EAV, lo que permite corroborar la elección del modelo semántico Mixto como la mejor opción para cumplir con los objetivos y las hipótesis planteadas en este trabajo de tesis. No obstante, ambos modelos semánticos solventan la hipótesis ***H1,*** porque permiten poblar una base de conocimientos financieros a partir de la integración de fuentes de datos externas.

Finalmente, con base en los resultados obtenidos en este análisis comparativo, los experimentos siguientes estarán centrados en el modelo semántico Mixto inspirado en Linked Data.

6.2 Descubrimiento y vinculación de datos financieros en la Web

La publicación constante de grandes volúmenes de información financiera por parte de diversas organizaciones del sector empresarial a través de sus Estados financieros es un hecho que puede ser explotado mediante el uso de tecnologías semánticas y Linked Data para la integración de sus datos (Radzimski et al., 2012). En este sentido, para validar la hipótesis ***H2*** es necesario ampliar las fronteras que el conjunto de datos financieros inspirado en Linked Data ofrece a través de sus datos. Para extender estas fronteras, se realizó un proceso que permite el descubrimiento de enlaces en la Web, propiamente en DBpedia, por ser el núcleo de la Linked Open Data cloud o LOD cloud (Auer et al., 2007; Bizer et al., 2009), nombre con el que ha sido referida esta nube de conjuntos de datos RDF. Este proceso incluye los pasos siguientes (Sánchez-Cervantes et al., 2013):

1. Elección del marco de trabajo para la ejecución de los experimentos.
2. Definición y ejecución de los experimentos para el descubrimiento de enlaces con y sin la aplicación de la distancia *Levenshtein* y *Stopwords* (Palabras vacías).
3. Análisis de los enlaces descubiertos mediante la aplicación de las métricas *Precision and Recall.*

Los pasos mencionados y los resultados obtenidos se describen en las secciones siguientes del presente apartado de tesis.

6.2.1 Elección del marco de trabajo para la ejecución de los experimentos para el descubrimiento de enlaces en la LOD cloud

Antes de iniciar con los experimentos para el descubrimiento de enlaces en DBpedia, es necesario seleccionar una herramienta que sirva de apoyo para este propósito. Para llevar a cabo esta selección, se analizaron los resultados obtenidos por Ngomo & Auer, (2011), que evaluaron los marcos de trabajo Silk y LIMES descritos en la sección 2.4.12 del Estado del arte. La evaluación realizada por estos autores, consistió en comparar el rendimiento de Silk y LIMES basándose en tres métricas: a) el tiempo necesario para obtener los datos de la base de conocimientos de origen y destino; b) el tiempo necesario para comparar las instancias y; c) el tiempo necesario para escribir los resultados *(output file)*. Sus experimentos se llevaron a cabo con tres configuraciones diferentes, el primer experimento, llamado *DrugBank*, consistió en mapear los medicamentos registrados en DBpedia y Linked Data DrugBank (Wishart et al., 2006) mediante la comparación de sus etiquetas, el objetivo del segundo experimento, llamado *SimCities*, fue detectar sitios duplicados dentro de ciudades registradas en DBpedia a través de la comparación de sus etiquetas, finalmente, el propósito del tercer experimento, al que nombraron *Diseases*, fue mapear las enfermedades registradas en MESH (HJ & G, 1994; Winnenburg & Bodenreider, 2014), contra sus homónimas contenidas en LinkedCT (Hassanzadeh et al., 2009) mediante la comparación de sus etiquetas. Para ambos marcos de trabajo, cada experimento fue ejecutado tres veces teniendo como resultado, que LIMES superó a Silk en todas las configuraciones experimentales previamente descritas.

Los resultados obtenidos en la evaluación realizada por Ngomo & Auer, (2011), ayudan a justificar la elección de LIMES como marco de trabajo para la realización de los experimentos que permitan el descubrimiento de enlaces contenidos en DBpedia. Como

consecuencia de esta justificación, en la sección siguiente se procede con la descripción y ejecución de estos experimentos.

6.2.2 Experimentos para el descubrimiento de enlaces con información financiera en la LOD cloud

Aunque el volumen de la información almacenada en el conjunto de datos financieros pueda incrementarse a través del tiempo, es importante investigar en qué medida la LOD cloud cubre el ámbito financiero y qué elementos contenidos en ella pueden estar relacionados con el mencionado conjunto de datos financieros.

El conjunto de datos financieros almacena un total de 11,864 empresas estadounidenses, esta cantidad de empresas junto con sus ratios financieros y valores, es la muestra total de información extraída y transformada en tripletas RDF hasta el tercer trimestre de 2014 (véase Capítulo 5) y es con la que se da inicio a los experimentos para el descubrimiento de enlaces en la LOD cloud a través de DBpedia.

Con el fin de crear un mapeo entre los conceptos del conjunto de datos financieros y DBpedia se consideraron las propiedades siguientes, a) *Label* (El nombre de la empresa); b) *Central Índex Key* (CIK) y c); *Ticker Simbol*. Sin embargo, mientras que el CIK y el *Ticker Simbol* conducirían a la obtención de mejores resultados, la mayoría de las empresas registradas en DBpedia carecen de estos datos, lo que deja a *Label* como la propiedad más viable para la realización del mapeo. Para facilitar el mapeo de conceptos, fue necesario realizar los siguientes dos pasos:

1. Se realizó una lista de *Stopwords* para el nombre de las empresas, incluyendo abreviaturas comerciales como "inc *(Incorporated)*", "ltd *(Limited)*" y "Co *(Company)*", por mencionar algunos.
2. Con base en las empresas registradas en el conjunto de datos financieros, se realizó una lista de sinónimos que incluye el nombre de la empresa y los *Stopwords* filtrados. Por ejemplo, la compañía "Apple Inc" tendría el sinónimo "Apple".

Para comparar las cadenas de caracteres *(Strings)*, se utilizó la distancia *Levenshtein*, que es una métrica que mide la diferencia entre dos secuencias. La distancia *Levenshtein* es el número mínimo de cambios de un sólo carácter para transformar una cadena de caracteres en otra. Las operaciones de edición para llevar a cabo esta transformación son inserciones, eliminaciones o sustituciones (Levenshtein, 1966).

Después de configurar el marco de trabajo LIMES y realizar los dos pasos previamente descritos, se ejecutaron los experimentos utilizando la propiedad *owl:sameAs* aplicada a las empresas contenidas en el conjunto de datos financieros contra el SPARQL endpoint de DBpedia *(dbpedia-owl:Company)*, con los siguientes parámetros:

- **Experimento 1:** utiliza la distancia *Levenshtein* con valor 0 y no utiliza *Stopwords*. Esto significa que no se utilizan sinónimos para el nombre de las empresas, sólo se usa su nombre oficial.
- **Experimento 2:** emplea la distancia *Levenshtein* con valor 0 y utiliza *Stopwords*.
- **Experimento 3:** aplica la distancia *Levenshtein* con valor 1 y no hace uso de *Stopwords* (como en el Experimento 1).
- **Experimento 4:** utiliza la distancia *Levenshtein* con valor 1, y se emplean *Stopwords*.

Los resultados obtenidos tras la ejecución de cada uno de los experimentos realizados se muestran en la Tabla 11.

EXPERIMENTOS PARA EL DESCUBRIMIENTO DE ENLACES EN LA LOD cloud			
Número de empresas en el conjunto de datos financieros: 11, 864			
No. Experimento	**Distancia *Levenshtein***	***Stopwords***	**Enlaces descubiertos**
1	0	No	96
2	0	Sí	1563
3	1	No	437
4	1	Sí	3652

Tabla 11. Resultados del descubrimiento de enlaces en la LOD cloud a través de DBpedia

Los resultados que se presentan en la Tabla 11, muestran como una simple comparación de cadenas de caracteres (Experimentos 1 y 3), proporciona un número reducido de enlaces asignados a las empresas registradas en el conjunto de datos financieros. Sin embargo, una diferencia notable se presenta con el aumento de la cantidad de enlaces descubiertos cuando se filtra la lista de sinónimos (Experimentos 2 y 4). Aunado a esto, los experimentos con valor 0 en la distancia *Levenshtein*, no requieren la realización de cambios para transformar el nombre oficial de la empresa en su análogo con los *Stopwords* filtrados, por el contrario, el valor 1 asignado a la distancia *Levenshtein* sí requiere la realización del cambio de 1 carácter para transformar el nombre oficial de la empresa a su equivalente con los *Stopwords* filtrados (Sánchez-Cervantes et al., 2013). Cabe mencionar que una distancia *Levenshtein* con valor superior a 1, disminuye la posibilidad de descubrir enlaces con información relacionada con las empresas almacenadas en el conjunto de datos

financieros, porque el número de cambios a realizar para la transformación de una cadena de caracteres en otra es mayor, incrementando el número de enlaces ambiguos.

Aunque los valores asignados para la distancia *Levenshtein* fueron 0 y 1, es importante mencionar que posiblemente la distancia *Levenshtein* entre las cadenas de caracteres incluya varios enlaces falsos (falsos positivos), especialmente para las empresas con nombre corto, pero, proporciona más alternativas posibles para el caso de las empresas con nombres largos, en los que la diferencia no es sólo una letra o un espacio en blanco.

Los experimentos ejecutados para el descubrimiento de enlaces en la LOD cloud proporcionan resultados que permiten estimar la cantidad de enlaces relacionados con el conjunto de datos financieros, sin embargo, es indispensable la ejecución de experimentos complementarios que ayuden a validar los enlaces descubiertos con la finalidad de identificar cuáles son los enlaces que verdaderamente contienen información relacionada con las empresas almacenadas en el conjunto de datos financieros. Estos experimentos requieren de un proceso de validación manual, el cual es descrito en la sección siguiente.

6.2.3 Validación de los enlaces con información financiera descubiertos en la LOD cloud

Los resultados de los experimentos para el descubrimiento de enlaces en la LOD cloud a través de DBpedia descritos en la sección anterior, muestran que la cobertura del ámbito financiero en la LOD cloud con respecto al conjunto de datos financieros cuya base es el modelo semántico Mixto, es bastante limitado. Aunado a esto, la falta de datos como el CIK o el *Ticker Simbol*, que podrían ser utilizados para identificar unívocamente conceptos de las empresas hace que la interconexión del conjunto de datos financieros con fuentes de información externa siga siendo una tarea difícil que requiere de diversas técnicas para la desambiguación y validación manual (Sánchez-Cervantes et al., 2013).

Con base en la información presentada en la Tabla 11, se eligió el resultado de los enlaces descubiertos en el Experimento 2 para acotar la cantidad de datos del conjunto de datos financieros con forme a las 500 empresas incluidas en el índice bursátil *Standard & Poor's 500* también conocido como índice S&P500[27] y de esa manera, realizar una validación de los enlaces descubiertos utilizando las métricas *Precision and Recall* (Precisión y Sensibilidad) que se describen más a delante.

[27]Standard & Poor's 500 (Datos hasta el 30 de Junio de 2014): http://us.spindices.com/indices/equity/sp-500

La elección del índice S&P500, se justifica porque incluye las empresas más representativas (las que tienen mayor número de ingresos) de la economía de Estados Unidos (Plerou et al., 1999), y porque la información almacenada en el conjunto de datos financieros corresponde a empresas del mencionado país. Así mismo, la elección de los enlaces descubiertos en el Experimento 2, se justifica por las métricas utilizadas para su obtención, que expresado en otras palabras, significa que aunque se filtró la lista de *Stopwords*, no fue necesaria la realización de cambios en los caracteres que conforman el nombre oficial de las empresas para transformarlos en su sinónimo, proporcionando así, enlaces de empresas que incluyen su nombre oficial junto con su abreviatura comercial, nombre con el que normalmente son registradas las empresas en documentos oficiales, como los Estados financieros.

El objetivo principal de validar manualmente los enlaces descubiertos, es eliminar las ambigüedades que puedan existir en ellos e identificar aquellos enlaces que realmente contienen información relacionada con la empresas almacenadas en el conjunto de datos financieros así como facilitar la validación de la hipótesis ***H2***. Esta validación consiste en emplear las métricas *Precision and Recall*, que son utilizadas en el dominio de la recuperación de información para medir qué tan bien un sistema busca, reconoce patrones y recupera la información (documentos) solicitada por un usuario. Siguiendo con este contexto, las métricas *Precision and Recall*, son definidas de la siguiente manera (Ting, 2010):

- *Precision* = Número total de documentos (enlaces) recuperados que son relevantes / número total de documentos (enlaces) que fueron recuperados.
- *Recall* = Número total de documentos (enlaces) recuperados que son relevantes / Número total de documentos relevantes en el conjunto de datos (Segmento de datos del índice S&P 500).

Estableciendo como referencia el segmento de 500 empresas incluidas en el índice S&P500 obtenidas a partir de los 1563 enlaces descubiertos en el Experimento 2, se aplicaron las métricas *Precision and Recall* y se obtuvieron los resultados que se muestran en la Tabla 12.

VALIDACIÓN DE LOS ENLACES DESCUBIERTOS EN LA LOD cloud	
Segmento de enlaces relevantes (S&P500): 500 **Número de enlaces recuperados en el conjunto de datos financieros:** 364 **Número de enlaces relevantes:** 330	
***Precision* (Precisión)**	0.91
***Recall* (Sensibilidad)**	0.66
Falsos positivos	34
Falsos negativos	170

Tabla 12. Resultados de la validación de los enlaces descubiertos

En un escenario en el que las métricas *Precision and Recall* son aplicadas, el resultado ideal es aquel en el que ambas métricas tienen un valor alto, es decir, un valor muy cercano a 1. En términos simples, la alta precisión significa que un algoritmo devuelve resultados sustancialmente más relevantes que irrelevantes, mientras que la alta sensibilidad significa que un algoritmo devuelve la mayor parte de los resultados relevantes (Carterette, 2009; Ting, 2010).

Los resultados que se presentan en la Tabla 12, indican que del segmento de 500 empresas que forman parte del índice S&P500, se recuperaron 364 enlaces correspondientes a empresas almacenadas en el conjunto de datos financieros, con una precisión relativamente alta del 91%, y con una sensibilidad media del 66%, esto significa que 330 enlaces son sustancialmente relevantes, es decir, que estos enlaces permiten vincular a 330 empresas del conjunto de datos financieros con información externa relacionada con ellas, en este caso DBpedia.

De los 364 enlaces obtenidos, 34 son falsos positivos y por lo tanto, no ofrecen información relevante para el conjunto de datos financieros. A diferencia de estos, los 170 enlaces falsos negativos, posiblemente proporcionarían información relevante para el conjunto de datos financieros, sin embargo, estos enlaces no fueron obtenidos durante la ejecución de los experimentos para el descubrimiento de enlaces en la LOD cloud a través de DBpedia (véase Sección 6.2.2), por lo que su análisis fue manual con la finalidad de incrementar la cantidad de enlaces descubiertos en DBpedia, cuya información este relacionada con sus respectivas empresas en el conjunto de datos financieros.

Es indispensable mencionar que aunque 330 enlaces descubiertos son sustancialmente relevantes, cada uno de ellos fue analizado manualmente para confirmar que las páginas Web en la LOD cloud con las que estos se vinculan contienen información fidedigna con respecto a las empresas del conjunto de datos financieros con las que estas páginas se

relacionan. De la misma manera, los enlaces falsos negativos fueron analizados directamente en DBpedia obteniendo un total de 69 enlaces que debieron ser descubiertos.

6.2.4 Conclusión del descubrimiento y vinculación de datos financieros en la Web

El descubrimiento de enlaces para vincular la información almacenada en el conjunto de datos financieros inspirado en Linked Data con información externa, en este caso la LOD cloud a través de DBpedia, es un proceso llevado a cabo meticulosamente, con el que analizando y sumando los enlaces relevantes y los enlaces falsos negativos, se obtuvo un total de 399 enlaces.

Los enlaces descubiertos y su vinculación con DBpedia, facilitan la validación de la hipótesis *H2* porque ayudan a comprobar que los datos procesados en la base de conocimientos financieros generada a partir del modelo semántico Mixto, permiten la reutilización de datos con terceros a través de Linked Data. Es importante mencionar que a través de DBpedia se ejemplificó la interconexión entre esta y el conjunto de datos financieros, sin embargo, este último podría vincularse con cualquier otra fuente de datos que pudiera considerarse relevante. Así mismo, no es objeto de esta tesis mejorar la tasa de aciertos en el descubrimiento automático de enlaces, sino dar soporte a dichos enlaces.

Un uso demostrativo del conjunto de datos financieros inspirado en Linked Data para confirmar la validación de la hipótesis ***H2*** y validar las hipótesis ***H3*** y ***H4***, se proporciona mediante el caso de estudio que se describe en la sección siguiente y posteriormente, a través del planteamiento de un conjunto de preguntas dirigidas hacia expertos en finanzas y contabilidad, cuyas respuestas permitirán constatar que la propuesta presentada a lo largo de este trabajo de tesis doctoral, supone o no, una mejora en la representación estructurada y unificada de datos financieros publicados por las empresas.

6.3 Caso de estudio: análisis comparativo de empresas por sector para la recomendación de una inversión

Como herramienta para la validación de las hipótesis ***H3*** y ***H4,*** se desarrolló un visualizador que permite acceder a los datos almacenados en el conjunto de datos financieros cuya base, es el modelo semántico Mixto inspirado en los principios de Linked Data, y que incluye heurísticas que sirven de apoyo para la toma de decisiones de manera automática.

Este visualizador, al que en adelante nombraremos FILIGRANT *(FInancial Linked Data GRaph ANAlysis Tool)* [28], permite ejemplificar un caso de estudio aplicado al conjunto de datos financieros, que consiste en analizar los datos financieros correspondientes a los años 2012, 2013 y hasta el tercer trimestre de 2014 de las empresas Wal-Mart Stores Inc. y Costco Wholesale Corp., pertenecientes al sector comercial. El objetivo de este caso de estudio es demostrar que con el uso de FILIGRANT, una institución bancaria a través de su departamento de Fondos de inversión puede recomendar a un cliente la conveniencia de invertir su dinero en alguna de las dos grandes empresas previamente mencionadas.

Una vez establecido el objetivo que se persigue con el caso de estudio, se plantea la siguiente situación:

1. Andrés Arroyo Ramírez desea invertir 200.000,00 USD en alguna empresa del sector comercial, sin embargo, Andrés desconoce la manera en la que puede tomar una decisión, por lo que acude al servicio de Fondos de inversión del banco BBVA Bancomer.
2. Los asesores financieros del departamento de Fondos de inversión del banco BBVA, hacen uso de FILIGRANT para analizar los ratios financieros de Wal-Mart y Costo que les permiten obtener las conclusiones necesarias para ofrecer a Andrés una recomendación para invertir.

De acuerdo con la situación planteada los asesores financieros del Fondo de inversión de BBVA realizan el análisis de los ratios de Liquidez de Wal-Mart y Costco con la finalidad de conocer la capacidad que dichas empresas tienen para convertir sus activos en caja, y medir la solvencia que estas poseen para cumplir con sus obligaciones a corto plazo. Además, en ambas empresas los asesores utilizan FILIGRANT con el fin de conocer las proporciones de las inversiones que han sido financiadas con recursos de terceros, es decir, el nivel de endeudamiento de las empresas. De esta manera, los asesores utilizan FILIGRANT para buscar, calcular, analizar, interpretar, concluir y recomendar a Andrés, la empresa en la que más le conviene invertir, fundamentándose en los resultados obtenidos entre el 01 de Enero de 2012 y el 11 de Noviembre de 2014, para los indicadores financieros adicionales siguientes (véase Tabla 2), Razón corriente, Capital de trabajo, Prueba ácida y Razón de deuda, que son descritos a detalle en los apartados siguientes.

[28] FILIGRANT: http://nadir.uc3m.es/flora-interface/

6.3.1 Razón corriente de las empresas Wal-Mart y Costco

Como se especificó en la Tabla 2 del Estado del arte, la Razón corriente es uno de los indicadores financieros que permite determinar el índice de liquidez de una empresa. El resultado obtenido es muy interesante, porque determina la capacidad de pago que tiene una empresa. Entre mayor sea la razón resultante, mayor solvencia y capacidad de pago tendrá la empresa, lo cual, es una garantía tanto para la empresa que no tendrá problemas para pagar sus deudas, como para sus acreedores, puesto que éstos tendrán certeza de que su inversión no se perderá (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010).

6.3.1.1 Razón corriente de la empresa Wal-Mart

La gráfica que se presenta en la Figura 23, muestra que dentro del periodo de búsqueda indicado, el tercer trimestre de 2014 ha sido el mejor periodo de Wal-Mart para solventar y pagar sus deudas vigentes. Sin embargo, su nivel de solvencia era bajo debido a que por cada Dólar que Wal-Mart tenía como deuda a corto plazo, disponía de 0.88 Centavos para saldarlo, dejando 0.12 Centavos como adeudo aún pendiente .

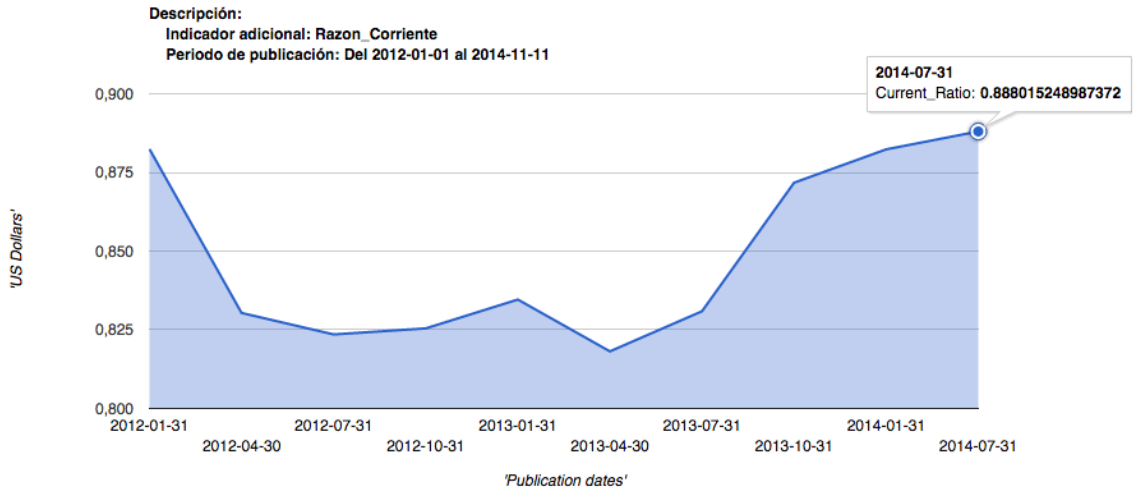


Figura 23. Gráfica trimestral de la Razón corriente para Wal-Mart

FILIGRANT, proporciona a los asesores financieros de BBVA un compendio de datos que facilita el análisis, cálculo e interpretación de la Razón corriente para la empresa Wal-Mart. Los datos que conforman este compendio, se muestra a detalle en la Figura 24, y corresponden a la grafica de la Figura 23.

*Valores en millones de Dólares.

Current_Assets	Current_Liabilities	Current_Ratio_Value	Publication_Date
54975000000	62300000000	0.882423756019262	2012-01-31
57276000000	68978000000	0.830351706341152	2012-04-30
56259000000	68315000000	0.823523384322623	2012-07-31
63431000000	76845000000	0.825440822434771	2012-10-31
59940000000	71818000000	0.834609707872678	2013-01-31
60176000000	73552000000	0.818142266695671	2013-04-30
60002000000	72214000000	0.830891516880383	2013-07-31
67142000000	77021000000	0.871736279715922	2013-10-31
61185000000	69345000000	0.882327492969933	2014-01-31
59632000000	67152000000	0.888015248987372	2014-07-31

Figura 24. Datos trimestrales de la Razón corriente para Wal-Mart

RAZÓN CORRIENTE POR AÑO DE WAL-MART DEL 01-01-2012 al 11-11-2014					
AÑO	**ACTIVO CORRIENTE**	**PASIVO CORRIENTE**	**RAZÓN CORRIENTE**	**OBLIGACIONES (%)**	**DISPONIBLE (%)**
2012	231.941.000.000,00	276.438.000.000,00	3.36 USD	83.90%	16.10%
2013	247.260.000.000,00	294.605.000.000,00	3.35 USD	83.93%	16.07%
2014	120.817.000.000,00	136.497.000.000,00	1.76 USD	88.51%	11.49%

Tabla 13. Razón corriente por año de la empresa Wal-Mart

Interpretación: al aplicar las siguientes reglas, ((Pasivo Corriente * 100) / Activo Corriente) y (((Activo Corriente – Pasivo Corriente) * 100) / Activo Corriente), sobre los datos anuales, se obtienen los porcentajes para el pago de obligaciones y capital disponible respectivamente. En este sentido, la Tabla 13, muestra que para el año 2012, por cada Dólar de obligación vigente (deuda) Wal-Mart contaba con 3.36 Dólares para respaldarla, esto significa que del 100% de sus ingresos (Activos Corrientes) el 83% eran para el pago de sus obligaciones y tan sólo el 16.10% le quedaron disponibles. Con respecto al año 2013, su liquidez disminuyó, el pago de sus obligaciones se incrementó al 83.93% y sus ingresos disponibles disminuyó al 16.07%. Finalmente, la Razón corriente del tercer trimestre de 2014 no indicó mejoría con relación a los años anteriores.

6.3.1.2 Razón corriente de la empresa Costco

La gráfica de la Figura 25, muestra que dentro del periodo de búsqueda del 01-01-2012 al 11-11-2014, al igual que Wal-Mart, el tercer trimestre de 2014 ha sido el mejor periodo de Costco para solventar y pagar sus deudas. Sin embargo, a diferencia de Wal-Mart, Costco contaba con un nivel de solvencia adecuado, ya que por cada Dólar que esta empresa tenia de deuda en corto plazo, disponía de 1.98 Dólares para solventarla.

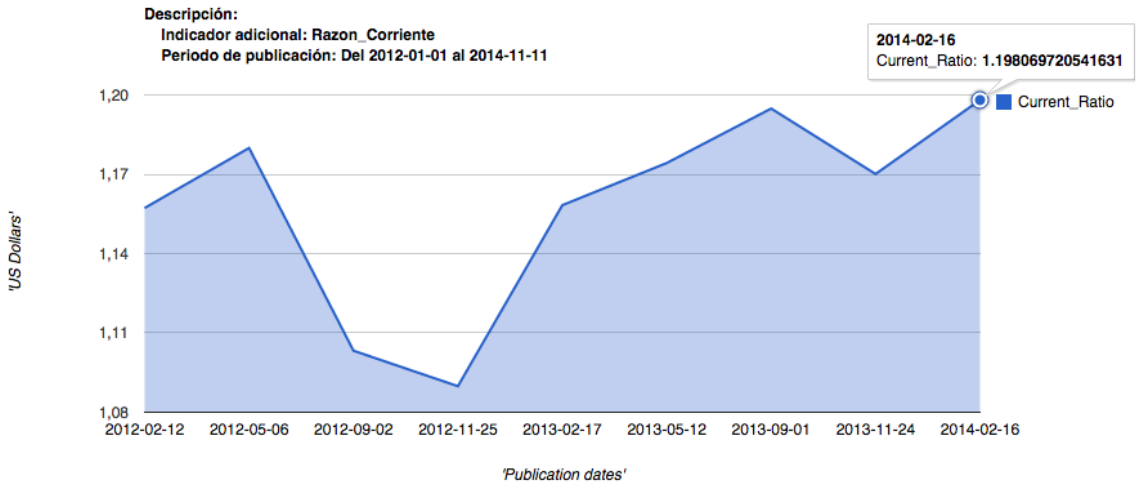


Figura 25. Gráfica trimestral de la Razón corriente para Costco

Al igual que con Wal-Mart, a través de FILIGRANT los asesores financieros de BBVA obtienen un resumen de datos que facilita el análisis, cálculo e interpretación de la Razón corriente para la empresa Costco. Estos datos se presentan detalladamente en la Figura 26 y corresponden a la grafica mostrada en la Figura 25.

*Valores en millones de Dólares.

Current_Assets	Current_Liabilities	Current_Ratio_Value	Publication_Date
14360000000	12409000000	1.157224595051978	2012-02-12
14591000000	12366000000	1.179928837134077	2012-05-06
13526000000	12260000000	1.10326264274062	2012-09-02
15405000000	14135000000	1.089847895295366	2012-11-25
15110000000	13045000000	1.158298198543503	2013-02-17
15863000000	13509000000	1.174254200903102	2013-05-12
15840000000	13257000000	1.194840461642906	2013-09-01
17473000000	14933000000	1.170093082434876	2013-11-24
16634000000	13884000000	1.198069720541631	2014-02-16

Figura 26. Datos trimestrales de la Razón corriente para Costco

RAZÓN CORRIENTE POR AÑO DE COSTCO DEL 01-01-2012 al 11-11-2014					
AÑO	ACTIVO CORRIENTE	PASIVO CORRIENTE	RAZÓN CORRIENTE	OBLIGACIONES (%)	DISPONIBLE (%)
2012	578.82.000.000,00	51.170.000.000,00	4.50 USD	113.12%	13.12%
2013	64.286.000.000,00	54.744.000.000,00	4.68 USD	117.43%	17.43%
2014	16.634.000.000,00	13.884.000.000,00	1.19 USD	119.81%	19.81%

Tabla 14. Razón corriente por año de la empresa Costco

Interpretación: tras aplicar las reglas utilizadas previamente en el análisis de la Razón corriente de Wal-Mart, en la Tabla 14, se interpretan los indicadores siguientes, para el año 2012, por cada Dólar de obligación vigente (deuda) Costco contaba con 4.50 Dólares para respaldarla, esto significa que del 100% de sus ingresos (Activos Corrientes) el 113.12%

eran para el pago de sus obligaciones, quedándole el 13.12% de sus ingresos disponibles. Con respecto al año 2013, tanto su liquidez para el pago de sus obligaciones como sus ingresos disponibles se incrementaron un 4.31%. Para terminar el análisis de la Razón corriente de Costco, se puede apreciar que para esta empresa, el conjunto de datos sólo contiene información financiera del primer trimestre del año 2014. La razón de esto, radica en la posibilidad de que la empresa Costo aún no haya registrado sus Estados financieros en el repositorio EDGAR, por lo cual, no han sido descargados ni transformados en tripletas RDF con base en el modelo semántico Mixto para que formen parte del conjunto de datos financieros inspirado en Linked Data que se utiliza en el presente capítulo para analizar la situación financiera de las empresas y principalmente, validar las hipótesis ***H2, H3*** y ***H4*** de este trabajo de tesis. Sin embargo, los datos que se tienen permiten determinar que hasta el primer trimestre del año 2014, Costco contaba con 1.19 Dólares para saldar cada Dólar que tuviese de adeudo vigente, por lo tanto, del 100% de sus ingresos en ese trimestre (Activos Corrientes), el 119% era para saldar sus deudas a corto plazo, y tan sólo el 19.81% de esos ingresos le quedaba disponible.

6.3.2 Capital de trabajo de las empresas Wal-Mart y Costco

El Capital de trabajo busca garantizar las operaciones de la empresa, si el resultado es positivo, da la posibilidad de generar inversión y si es negativo, da la posibilidad de buscar financiamiento ya se a propio o mediante fondos de terceros (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010). En las Figuras 27 y 28, se especifican los datos obtenidos por FILIGRANT para facilitar a los asesores financieros de BBVA, el análisis, cálculo e interpretación por año del Capital de trabajo de Wal-Mart y Costco respectivamente.

*Valores en millones de Dólares.

Current_Assets	Current_Liabilities	Working_Capital_Value	Publication_Date
54975000000	62300000000	-7325000000	2012-01-31
57276000000	68978000000	-11702000000	2012-04-30
56259000000	68315000000	-12056000000	2012-07-31
63431000000	76845000000	-13414000000	2012-10-31
59940000000	71818000000	-11878000000	2013-01-31
60176000000	73552000000	-13376000000	2013-04-30
60002000000	72214000000	-12212000000	2013-07-31
67142000000	77021000000	-9879000000	2013-10-31
61185000000	69345000000	-8160000000	2014-01-31
59632000000	67152000000	-7520000000	2014-07-31

Figura 27. Datos trimestrales del Capital de trabajo para Wal-Mart

Los resultados especificados en la Figura 27, sugieren que Wal-Mart no cuenta con la liquidez suficiente para pagar sus deudas a corto plazo, sin embargo, esto no significa que la empresa esté en quiebra o que haya suspendido sus pagos.

*Valores en millones de Dólares.

Current_Assets	Current_Liabilities	Working_Capital_Value	Publication_Date
14360000000	12409000000	1951000000	2012-02-12
14591000000	12366000000	2225000000	2012-05-06
13526000000	12260000000	1266000000	2012-09-02
15405000000	14135000000	1270000000	2012-11-25
15110000000	13045000000	2065000000	2013-02-17
15863000000	13509000000	2354000000	2013-05-12
15840000000	13257000000	2583000000	2013-09-01
17473000000	14933000000	2540000000	2013-11-24
16634000000	13884000000	2750000000	2014-02-16

Figura 28. Datos trimestrales del Capital de trabajo para Costco

A diferencia de Wal-Mart, la Figura 28 muestra que Costco tiene un Capital de trabajo positivo lo que se traduce en que la empresa tiene la capacidad de satisfacer sus obligaciones (deudas) a corto plazo con la simple recuperación o liquidación de sus Activos Corrientes.

Con la finalidad de simplificar los resultados obtenidos para el Capital de trabajo de ambas empresas, en las Tablas 15 y 16 estos son presentados de manera anual.

CAPITAL DE TRABAJO POR AÑO DE WAL-MART DEL 01-01-2012 al 11-11-2014			
AÑO	**ACTIVO CORRIENTE**	**PASIVO CORRIENTE**	**CAPITAL DE TRABAJO**
2012	231.941.000.000,00	276.438.000.000,00	-44.497.000.000,00
2013	197.260.000.000,00	294.605.000.000,00	-47.345.000.000,00
2014	120.817.000.000,00	136.497.000.000,00	-15.680.000.000,00

Tabla 15. Capital de trabajo por año de la empresa Wal-Mart

Interpretación: como ya se ha mencionado, los resultados negativos en el Capital de trabajo de Wal-Mart (véase Tabla 15) reflejan inestabilidad en su liquidez para el pago de sus deudas a corto plazo, sin que esto signifique que la empresa se encuentre en quiebra o haya suspendido el pago de sus obligaciones; no obstante, estos resultados sugieren la necesidad de buscar financiamiento ya sea propio o mediante fondos de terceros que permitan incrementar los Activos Corrientes y que el Capital de trabajo pase de negativo a positivo, permitiendo el pago de deudas e intereses así como la generación de utilidad.

CAPITAL DE TRABAJO POR AÑO DE COSTCO DEL 01-01-2012 al 11-11-2014			
AÑO	**ACTIVO CORRIENTE**	**PASIVO CORRIENTE**	**CAPITAL DE TRABAJO**
2012	57.882.000.000,00	51.170.000.000,00	6.712.000.000,00
2013	64.286.000.000,00	54.744.000.000,00	9.542.000.000,00
2014	16.634.000.000,00	13.884.000.000,00	2.750.000.000,00

Tabla 16. Capital de trabajo por año de la empresa Costco

Interpretación: en comparación con Wal-Mart, para el año 2012, una vez que Costco canceló sus obligaciones, le quedaron 6.712.000.000,00 USD para atender las obligaciones que surgen en el normal desarrollo de su actividad económica. Además, se puede observar que para el año 2013, su Capital de trabajo se incrementó 2.830.000.000,00 USD. Este incremento puede ser consecuencia de un plan de inversiones ejecutado por la compañía. En lo que respecta al Capital de trabajo correspondiente al año 2014, únicamente se muestran datos hasta el primer trimestre de ese año, indicando un incremento de 799,000,000,00 USD en comparación con el primer trimestre del año 2012 y un incremento de 685,000,000,00 USD en comparación con el primer trimestre del año 2013 (para ambos casos, véase Figura 28).

En lo que respecta al apoyo para la toma de decisiones de manera automática, FILIGRANT compara a través de sus heurísticas, los resultados trimestrales del Capital de trabajo de Wal-Mart y Costco. Al realizar una comparación entre ambas empresas, FILIGRANT identifica que Wal-Mart posee valores negativos, por lo que realiza las dos recomendaciones siguientes:

1. Sería recomendable obtener más financiamiento para Wal-Mart.
2. En comparación con Wal-Mart, la empresa Costco podría ser la mejor opción para invertir.

6.3.3 Prueba ácida de las empresas Wal-Mart y Costco

La Prueba ácida revela la capacidad de la empresa para cancelar sus obligaciones corrientes pero sin contar con la venta de sus existencias. Si es menor a 1, la empresa podría suspender sus pagos con terceros y si es mayor que 1 indica la posibilidad de que haya un exceso de liquidez, lo que sería ideal para la empresa. Los valores óptimos se encuentran entre 0,5 y 1 (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010). La Figura 29, muestra los datos obtenidos por FILIGRANT para facilitar a los asesores financieros de BBVA el análisis, cálculo e interpretación por año del indicador de Prueba ácida correspondiente a la empresa Wal-Mart.

*Valores en millones de Dólares.

Current_Assets	Current_Liabilities	Inventory	AcidTest_Value	publication_Date
54975000000	62300000000	40714000000	0.228908507223114	2012-01-31
57276000000	68978000000	41284000000	0.231842036591377	2012-04-30
56259000000	68315000000	40558000000	0.229832394056942	2012-07-31
63431000000	76845000000	47487000000	0.207482594833756	2012-10-31
59940000000	71818000000	43803000000	0.224692973906263	2013-01-31
60176000000	73552000000	43138000000	0.231645638459865	2013-04-30
60002000000	72214000000	42793000000	0.238305591713518	2013-07-31
67142000000	77021000000	49673000000	0.226808273068384	2013-10-31
61185000000	69345000000	44858000000	0.235445958612733	2014-01-31
59632000000	67152000000	45451000000	0.21117762687634	2014-07-31

Figura 29. Datos trimestrales de la Prueba ácida para Wal-Mart

Interpretación: con base en los resultados de la Prueba ácida *(AcidTest_Value)* aplicada a la empresa Wal-Mart (véase Figura 29), se observa que durante los trimestres comprendidos entre el 01 de Enero de 2012 y el 11 de Noviembre de 2014, la empresa mantuvo valores de entre 0.21 y 0.22, los cuales, están por debajo del valor óptimo indicado para esta prueba. Si en algún momento la empresa tuvo la necesidad de atender todas sus obligaciones corrientes sin la necesidad de liquidar y vender sus inventarios, a esta, no le habría alcanzado y por lo tanto, habría tenido que vender sus inventarios (o parte de ellos) para poder cumplir con sus obligaciones. Analizando el resultado de Prueba ácida más reciente de los presentados en la Figura 29, se puede deducir que la empresa disponía solamente del 21% de sus activos líquidos para cubrir sus deudas a corto plazo, que expresado de otra manera, significa que sus deudas a corto plazo son un 79% superiores a sus activos líquidos.

Continuando con el análisis del indicador de Prueba ácida, en la Figura 30, se proporcionan los datos financieros obtenidos por FILIGRANT para facilitar el análisis, cálculo e interpretación por año de esta prueba aplicada sobre la empresa Costco.

*Valores en millones de Dólares.

Current_Assets	Current_Liabilities	Inventory	AcidTest_Value	publication_Date
14360000000	12409000000	6934000000	0.598436618583286	2012-02-12
14591000000	12366000000	7044000000	0.610302442180171	2012-05-06
13526000000	12260000000	7096000000	0.524469820554649	2012-09-02
15405000000	14135000000	8152000000	0.513123452423063	2012-11-25
15110000000	13045000000	7582000000	0.57707934074358	2013-02-17
15863000000	13509000000	7635000000	0.609075431194019	2013-05-12
15840000000	13257000000	7894000000	0.599381458851927	2013-09-01
17473000000	14933000000	9337000000	0.544833590035492	2013-11-24
16634000000	13884000000	8267000000	0.602636127917027	2014-02-16

Figura 30. Datos trimestrales de la Prueba ácida para Costco

Interpretación: aplicando el mismo criterio de análisis utilizado con Wal-Mart, la empresa Costco mantuvo valores de entre 0.51 y 0.61 para su Prueba ácida, que, a diferencia de Wal-Mart, estos valores sí se encuentran dentro del rango considerado como óptimo para esta prueba. Basándose en el último de los resultados de Prueba ácida obtenidos en la Figura 30, se observa que Costco disponía del 60% de sus activos líquidos para saldar sus deudas a corto plazo, que descrito en otros términos, significa que sus deudas a corto plazo sólo son un 40% superiores a sus activos líquidos. Al igual que Wal-Mart, si en algún momento Costco tuvo la necesidad de atender todas sus obligaciones corrientes sin la necesidad de liquidar y vender sus inventarios, a esta, no le habría alcanzado y por lo tanto, habría tenido que vender sus inventarios (o parte de ellos) para poder cumplir con sus obligaciones.

Como apoyo a la toma de decisiones de forma automática, FILIGRANT a través de sus heurísticas compara los resultados trimestrales obtenidos para la Prueba ácida aplicada a Wal-Mart y Costco e identifica que los valores obtenidos para Wal-Mart se encuentran por debajo del valor óptimo sugerido para esta prueba, mientras que también identifica que los valores de Prueba ácida de Costco se encuentran dentro del valor óptimo sugerido. Al realizar una comparación, FILIGRANT sugiere invertir en Costco.

6.3.4 Razón de deuda para las empresas Wal-Mart y Costco

La Razón de deuda permite establecer el grado de participación de los acreedores, en los activos de la empresa. Mide la proporción de la inversión de la empresa que ha sido financiada por deuda, por lo que se acostumbra presentar su resultado en forma de porcentaje. Se considera que un endeudamiento del 60% es manejable (Kimmel et al., 2010; Montero & Fernández-Aviles, 2010). Por ejemplo, de cada 100.00 USD que una empresa tiene en sus activos se adeudan 60.00 USD. Esta razón de endeudamiento indica que el 60% del total de la inversión (Activos Totales) ha sido financiada con recursos de terceros (endeudamiento).

Para el análisis de la Razón de Endeudamiento de Wal-Mart y Costo, los asesores financieros del Fondo de inversiones de BBVA requieren aplicar la fórmula siguiente (véase Tabla 3):

$$\text{Razón de deuda} = ((\text{Pasivos totales} / \text{Activos Totales}) * 100)$$

Para la aplicación de esta fórmula, es necesario obtener los ratios financieros que conforman a los Pasivos Totales y a los Activos Totales, los cuales son los siguientes:

a) Pasivos Totales (Pasivos Corrientes + Pasivos No Corrientes)

a. Pasivos Corrientes *(Liabilities Current).*

b. Deuda a largo plazo, vencimientos actuales totales, *(Long-term Debt, Current Maturities, Total).*

b) Activos Totales (Activos Corrientes + Activos Fijos)

a. Activos Corrientes *(Current Assets).*

b. Depreciación acumulada, agotamiento y amortización, de propiedades, planta y equipo, *(Accumulated Depreciation, Depletion and Amortization, Property, Plant, and Equipment).*

Para las empresas analizadas, los ratios financieros listados previamente son buscados por los asesores de BBVA mediante FILIGRANT, para que posteriormente, estos apliquen la fórmula de Razón de deuda y realicen las interpretaciones pertinentes. Con relación a lo anterior, la información concerniente a la Razón de deuda correspondiente a las empresas Wal-Mart y Costco se describe en los apartados siguientes.

6.3.4.1 Razón de deuda para la empresa Wal-Mart

Las Figura 31, proporciona los datos trimestrales de los Pasivos Corrientes de la empresa Wal-Mart, del mismo modo, en la Figura 32 se muestran los datos referentes a las Deudas a largo plazo, en las que se incluyen los vencimientos actuales totales de la mencionada empresa. Ambos indicadores financieros sumados conforman los Pasivos totales de Wal-Mart. Adicionalmente, las Figura 33 y 34, muestran los Activos Corrientes y la Depreciación acumulada, agotamiento y amortización de propiedades, planta y equipo de Wal-Mart. Los dos indicadores financieros sumados, conforman los Activos totales de esta empresa.

*Valores en millones de Dólares.

Value_Ratio	Publication_date
62300000000	2012-01-31
68978000000	2012-04-30
68315000000	2012-07-31
76845000000	2012-10-31
71818000000	2013-01-31
73552000000	2013-04-30
72214000000	2013-07-31
77021000000	2013-10-31
69345000000	2014-01-31
67152000000	2014-07-31

Figura 31. Pasivos Corrientes de Wal-Mart

Value_Ratio	Publication_date
1975000000	2012-01-31
2509000000	2012-04-30
4029000000	2012-07-31
6550000000	2012-10-31
5587000000	2013-01-31
5967000000	2013-04-30
4692000000	2013-07-31
4147000000	2013-10-31
4103000000	2014-01-31
4659000000	2014-07-31

Figura 32. Deudas a largo plazo Wal-Mart

Value_Ratio	Publication_date
54975000000	2012-01-31
57276000000	2012-04-30
56259000000	2012-07-31
63431000000	2012-10-31
59940000000	2013-01-31
60176000000	2013-04-30
60002000000	2013-07-31
67142000000	2013-10-31
61185000000	2014-01-31
59632000000	2014-07-31

Figura 33. Activos Corrientes de Wal-Mart

*Valores en millones de Dólares.

Value_Ratio	Publication_date
45399000000	2012-01-31
47600000000	2012-04-30
48961000000	2012-07-31
50450000000	2012-10-31
51896000000	2013-01-31
53395000000	2013-04-30
54724000000	2013-07-31
56313000000	2013-10-31
57725000000	2014-01-31
61709000000	2014-07-31

Figura 34. Depreciación acumulada, agotamiento y amortización de propiedades, planta y equipo de Wal-Mart

Tras la obtención de los Pasivos totales y los Activos totales de Wal-Mart a través de FILIGRANT, los asesores del Fondo de inversiones de BBVA calculan la Razón de deuda de la citada empresa, de esta manera, analizan e interpretan los resultados obtenidos y mostrados por año en la Tabla 17.

RAZÓN DE DEUDA POR AÑO DE WAL-MART DEL 01-01-2012 al 11-11-2014			
AÑO	**PASIVOS TOTALES**	**ACTIVOS TOTALES**	**RAZÓN DE DEUDA (%)**
2012	291501000000,00	424351000000,00	68%
2013	314998000000,00	463588000000,00	67%
2014	145259000000,00	240251000000,00	60%

Tabla 17. Razón de deuda por año de la empresa Wal-Mart

Interpretación: los resultados que se presentan en la Tabla17, indican que para el año 2012, el 68% del total de la inversión (Activos Totales) de Wal-Mart fue financiado con la participación de sus acreedores, lo que representaba un moderado nivel de riesgo para la empresa, ya que por cada 100,00 USD que Wal-Mart tenia en sus Activos Totales, adeudaba 68,00 USD. Esta situación disminuyo un 1% en 2013, lo que significa que su nivel de endeudamiento seguía siendo arriesgado, sin embargo, hasta el tercer trimestre de 2014, su nivel de deuda indicó una mejoría con una disminución hasta el 60%, lo que se considera como un nivel de endeudamiento manejable.

6.3.4.2 Razón de deuda para la empresa Costco

Al igual que con Wal-Mart, en las Figuras 35 y 36, se proporcionan los indicadores trimestrales que sumados conforman los Pasivos totales de la empresa Costco.

Value_Ratio	Publication_date
12409000000	2012-02-12
12366000000	2012-05-06
12260000000	2012-09-02
14135000000	2012-11-25
13045000000	2013-02-17
13509000000	2013-05-12
13257000000	2013-09-01
14933000000	2013-11-24
13884000000	2014-02-16

Figura 35. Pasivos Corrientes de Costco

*Valores en millones de Dólares.

Value_Ratio	Publication_date
900000000	2012-02-12
0	2012-05-06
1000000	2012-09-02
1000000	2012-11-25
1000000	2013-02-17
0	2013-05-12

Figura 36. Deudas a largo plazo Costco

Por otro lado, en las Figura 37 y 38, se proporcionan los indicadores financieros, conforma los Activos totales de Costco.

Value_Ratio	Publication_date
14360000000	2012-02-12
14591000000	2012-05-06
13526000000	2012-09-02
15405000000	2012-11-25
15110000000	2013-02-17
15863000000	2013-05-12
15840000000	2013-09-01
17473000000	2013-11-24
16634000000	2014-02-16

Figura 37. Activos Corrientes de Costco

*Valores en millones de Dólares.

Value_Ratio	Publication_date
6268000000	2012-02-12
6443000000	2012-05-06
6585000000	2012-09-02
6772000000	2012-11-25
6933000000	2013-02-17
7091000000	2013-05-12
7141000000	2013-09-01
7333000000	2013-11-24
7486000000	2014-02-16

Figura 38. Depreciación acumulada, agotamiento y amortización de propiedades, planta y equipo de Costco

Del mismo modo que con Wal-Mart, después de obtener los Pasivos totales y los Activos totales de Costco a través de FILIGRANT, los asesores del Fondo de inversiones de BBVA calculan la Razón de deuda de la nombrada empresa, de este modo, analizan e interpretan los resultados que se muestran en la Tabla 18.

RAZÓN DE DEUDA POR AÑO DE COSTCO DEL 01-01-2012 al 11-11-2014			
AÑO	**PASIVOS TOTALES**	**ACTIVOS TOTALES**	**RAZÓN DE DEUDA (%)**
2012	52072000000,00	83950000000,00	62%
2013	54745000000,00	92784000000,00	59%
2014	13884000000,00	24120000000,00	57%

Tabla 18. Razón de deuda por año de la empresa Costco

Interpretación: los resultados que se presentan en la Tabla 18, permiten observar que los niveles de endeudamiento de Costco en comparación con el de Wal-Mart son menores y por lo tanto manejables. Particularmente para el año 2013 y hasta el tercer trimestre de 2014, Costo se encontraba con la capacidad de contraer más obligaciones o deuda.

6.3.5 Obtención de información adicional de las empresas

Posterior al análisis de los indicadores financieros, los asesores del Fondo de inversiones de BBVA deciden buscar información complementaria acerca de Wal-Mart y Costco a través de FILIGRANT.

La Figura 39, muestra un fragmento de las interfaces de DBpedia que contienen información de Wal-Mart y Costo y que son accedidas a través de FILIGRANT.

Figura 39. Enlaces a DPpedia a través del visualizador del conjunto de datos financieros

Mediante los enlaces a DBpedia, los asesores del Fondo de inversión de BBVA obtienen información adicional acerca de Wal-Mart y Costco que les permite complementar la recomendación que estos proporcionan a Andrés para la realización de su inversión. Estos enlaces forman parte de los resultados obtenidos en los experimentos para el descubrimiento de enlaces con información financiera en la LOD cloud descritos en la sección 6.2.2 del presente capítulo.

Después de todo un proceso de búsquedas, cálculos, análisis, interpretación de resultados y navegación en la Web de DBpedia para obtener información complementaria, los asesores financieros de BBVA proporcionan a Andrés las conclusiones que se describen en el apartado siguiente.

6.3.6 Conclusiones del Caso de estudio

Al realizar un análisis clásico de los indicadores financieros de las empresas Wal-Mart y Costco utilizando el conjunto de datos financieros inspirado en Linked Data y su visualizador para el periodo comprendido entre el 01 de Enero de 2012 y el 11 de Noviembre de 2014, los asesores financieros obtuvieron las conclusiones siguientes:

- Tanto Wal-Mart como Costco son empresas del sector comercial que presentan alta dependencia de las ventas de sus inventarios para cumplir con el pago de sus obligaciones corrientes (si le fueran exigidas) a corto plazo. Hasta este punto, no

hay diferencia entre las empresas y ambas son candidatas para la inversión de 200.000,00 USD que pretende realizar Andrés.

- La interpretación de la Razón corriente aplicada a Wal-Mart y Costo aparenta un equilibrio en la capacidad de pago que tienen ambas empresas, sin embargo, para ser precisos, la Razón corriente de Costco supera a la de Wal-Mart con 1.14 y 1.33 Dólares para los años 2012 y 2013 respectivamente. Mientras que Walt-Mar sólo supera a Costco con 0.57 Centavos en el primer trimestre de 2014. Esto permite inferir que Costco presenta mejor liquidez para saldar sus deudas a corto plazo. Es importante mencionar que en ambas empresas no se realizó la comparación de los trimestres 2 y 3 porque no se tenia la información financiera de Costco en esos periodos de tiempo.
- En lo que respecta al Capital de trabajo, este también favorece a Costco pues a diferencia de Wal-Mart, Costco no presenta resultados negativos, posiblemente por la aplicación de un adecuado plan de inversiones. Walt-Mar por su parte, requiere buscar financiamiento ya sea propio a través de terceros.
- Los resultados obtenidos de la Prueba ácida aplicada en ambas empresas, indican que Costco mantuvo los valores considerados óptimos (entre 0.51 y 0.61) para esta prueba. Mientras que Walt-Mar se mantuvo entre 0.21 y 0.22, lo que significa que si en algún momento esta empresa requirió realizar el pago de todas sus obligaciones corrientes sin la necesidad de liquidar y vender sus inventarios, a esta, no le habría alcanzado.
- Los resultados obtenidos tras el cálculo de la Razón de deuda, demuestran que Walt-Mar en los años 2012 y 2013, no mantuvo un nivel manejable en este ratio, con 68% y 67% de endeudamiento para cada uno de estos años. Hasta el tercer trimestre de año 2014, Wal-Mart presentó un nivel de endeudamiento aceptable con 60%. Sin embargo, Costco en el mismo periodo de tiempo, mantuvo un nivel de endeudamiento manejable con 62%, 59% y 57% permitiéndole la posibilidad de contraer más obligaciones o deudas.
- Después de analizar los ratios de liquidez y endeudamiento de las empresas Walt-Mar y Costco, los asesores recomiendan a Andrés invertir en la empresa Costco Wholesale Corp. El motivo de esta decisión se fundamenta en la liquidez que esta empresa tiene para saldar sus deudas a corto plazo, así como en su disponibilidad para contraer obligaciones.

El propósito del caso de estudio descrito en esta sección es demostrar que el modelo semántico Mixto inspirado en los principios de Linked Data presentado a lo largo de este trabajo de tesis, sirve de apoyo para facilitar la toma de decisiones utilizando como herramienta a FILIGRANT.

Los resultados obtenidos en este caso de estudio permiten validar las hipótesis ***H3*** y ***H4***. La hipótesis ***H3*** es validada porque a través del uso de tecnologías semánticas los datos financieros son transformados y estructurados en un formato semántico (RDF) para su publicación en la Web, tomando como modelo de datos, modelo semántico Mixto y las taxonomías financieras de Hoja de balance, Estado de flujo de efectivo y Cuenta de estado de resultados que lo integran.

Adicionalmente, la validación de la hipótesis ***H3*** es comprobada mediante las respuestas proporcionadas por 10 expertos en finanzas y contabilidad a través del cuestionario publicado en la Web de FILIGRANT. En sus respuestas, estos expertos indican que la iniciativa presentada en este trabajo de tesis, es una mejora cualitativa en la representación de los datos financieros que se publican en la Web.

En lo referente a la hipótesis ***H4***, esta es verificada de dos maneras, la primera consiste en validar el apoyo a la toma de decisiones por parte de las personas. Esta validación se fundamenta en que los asesores financieros de BBVA con base en su experiencia calculan, analizan e interpretan los resultados obtenidos con FILIGRANT para los indicadores financieros de Razón corriente, Capital de trabajo, Prueba ácida y Razón de deuda, aplicados a las empresas Wal-Mart y Costco, de este modo, recomiendan a su cliente Andrés Arrollo Ramírez invertir su dinero en la empresa Costco.

La segunda validación de la hipótesis *H4*, concierne al apoyo para la toma de decisiones de manera automática. En este sentido, las recomendaciones automáticas proporcionadas por FILIGRANT se realizan a través de sus heurísticas, mediante las que esta herramienta realiza una comparación de los resultados obtenidos para los ratios financieros del Capital de trabajo y la Prueba ácida aplicados a las empresas Wal-Mart y Costco. La manera en la que FILIGRANT proporciona estas recomendaciones, se describe en los dos párrafos siguientes.

Para el Capital de trabajo de ambas empresas, FILIGRANT compara los resultados trimestrales obtenidos con base en las dos reglas siguientes: a)Identificar cuál empresa es la que tiene los resultados más bajos y; b)Identificar si alguna de las dos empresas tiene

valores negativos. En este sentido, FILIGRANT identificó que los valores del Capital de trabajo de Wal-Mart son negativos, por lo que de manera automática realizó las siguientes recomendaciones: a) Se recomienda obtener mayor financiamiento para Wal-Mart y; b) En comparación con Wal-Mart, Costco es la empresa más viable para realizar una inversión.

Respecto a la validación de la hipótesis ***H4*** mediante la Prueba ácida aplicada a Wal-Mart y Costco, FILIGRANT compara los valores trimestrales obtenidos para ambas empresas considerando la siguiente regla, si el valor de Prueba ácida es inferior a 1, se recomienda no invertir en la empresa porque tiene problemas de liquidez, si el valor de Prueba ácida es mayor que 1 o, se encuentra entre 0,5 y 1, se recomienda invertir en la empresa porque cuenta con la liquidez suficiente para cumplir con sus obligaciones a corto plazo. En este aspecto, FILIGRANT después de comparar los valores de Prueba ácida de Wal-Mart y Costco, de manera automática proporciona la siguiente recomendación a los asesores financieros, se recomienda invertir en Costo porque el valor de su Prueba ácida se encuentra entre los valores óptimos para esta prueba.

Es importante resaltar, que las gráficas, tablas y valores obtenidos en el caso de estudio, han sido generados por FILIGRANT, todos los resultados surgen por la aplicación de esta herramienta durante las distintas etapas del proceso de recomendación. Adicionalmente, los cálculos, el análisis, interpretaciones y conclusiones proporcionadas en el caso de estudio, han sido verificados por un experto en contabilidad y análisis financiero.

Con el fin de complementar y corroborar las validaciones llevadas a cabo sobre el modelo semántico Mixto inspirado en Linked Data que permiten comprobar las hipótesis planteadas en este trabajo de tesis, es importante conocer la opinión de expertos en finanzas y contabilidad. Por este motivo, en la sección siguiente se describen las opiniones y sugerencias de un grupo de expertos en estas áreas del conocimiento a través de un cuestionario.

6.4 Validación de resultados a través de expertos

Para la validación de los resultados a través de expertos, se realizó un cuestionario conformado por 10 preguntas, este cuestionario está a disposición de los usuarios a través de la página Web de FILIGRANT.

El primer punto a considerar para la realización del cuestionario que se presenta en esta sección, es validar las hipótesis ***H3*** y ***H4***, además de constatar la utilidad que tiene el modelo semántico Mixto inspirado en los principios de Linked Data a través de

FILIGRANT para los conocedores en el campo de las finanzas y la contabilidad. Así mismo, se pretende observar si existen diferencias entre las respuestas y opiniones proporcionadas por los expertos a las preguntas planteadas en el cuestionario, con la finalidad de que en el trabajo a futuro de este trabajo de tesis, se apliquen las mejoras necesarias con base a las recomendaciones proporcionadas por los expertos.

Las preguntas que conforman el cuestionario realizado son descritas a continuación:

Pregunta 1.

¿Considera que los ratios financieros presentados para cada empresa son necesarios para analizar el estado de una compañía?

Pregunta 2.

¿Considera necesario añadir algún otro ratio financiero?

Pregunta 3.

En caso de que su respuesta haya sido sí, por favor indique ¿Cuáles ratios financieros?

Pregunta 4.

¿Considera que los ratios financieros incluidos facilitan el cálculo de otros indicadores financieros de las compañías?

Pregunta 5.

Los ratios y los indicadores financieros adicionales presentados, ¿Son útiles para realizar un análisis clásico de la situación financiera de las empresas?

Pregunta 6.

Como ejemplo demostrativo, se proporcionan enlaces a DBpedia, sin embargo, los datos pueden ser enlazados a otras fuentes de información financiera así como noticias, gobierno y educación, por mencionar algunas. ¿Cree que es útil vincular la información financiera de las empresas con información adicional relacionada con ellas procedente de otras fuentes de datos?

Pregunta 7.

¿Cree que presentar la información financiera de forma unificada, con indicadores adicionales y complementándola con información de otras fuentes de datos, puede facilitar la toma decisiones?

Pregunta 8.

Existen diversos formatos de publicación de información financiera. ¿Le parece necesario ofrecer este tipo de información en un formato unificado, conocido por todas las partes, de forma que todos utilicen un modelo similar y pueda ser reutilizado?

Pregunta 9.

En este sentido, ¿considera que la propuesta que le hemos presentado supone una mejora cualitativa en esta representación?

Pregunta 10.

Si las tiene, por favor indique sus opiniones y recomendaciones.

Los resultados obtenidos después de la realización de la encuesta, son proporcionados en el apartado siguiente.

6.4.1 Resultados y discusión

El análisis y discusión descritos en este apartado de tesis, corresponden a las respuestas proporcionadas por 10 encuestados. En la Tabla 19, se presenta un compendio de resultados que facilita la descripción de las deducciones que se describen a continuación.

RESPUESTAS A PREGUNTAS DEL CUESTIONARIO			
PREGUNTAS	**RESPUESTAS**		**RATIOS SUGERIDOS (PREGUNTAS 2, 3)**
	SÍ	**NO**	
1	9	1	No hubo ratios Sugeridos por parte de los encuestados.
2	10	0	
3	10	0	
4	10	0	
5	10	0	
6	10	0	
7	10	0	
8	10	0	
9	10	0	

Tabla 19. Respuestas de expertos a preguntas del cuestionario

Las respuestas de la pregunta 1, indican que el 90% de los encuestados considera que los ratios financieros incluidos en FILIGRANT, son necesarios para el análisis financiero de una empresa sin embargo, el 10% es decir, 1 encuestado indicó que estos ratios no son necesarios para conocer la situación financiera de una empresa. Con relación a esto, al analizar a detalle las opiniones y recomendaciones proporcionadas por los expertos (véase

Tabla 20), el doctorando se percató que esta respuesta corresponde al encuestado 1 y en su recomendación menciona que si estos ratios financieros forman parte de una fórmula que se derive de ellos, entonces sí podrían ayudar a conocer el estado financiero de las empresas. En lo que concierne a la pregunta 2, el 100% de los encuestados coinciden en que no es necesario añadir más ratios financieros y por ende, no se indicaron ratios para la pregunta número 3.

Continuando con el análisis de las respuestas proporcionadas por los expertos, en las preguntas 4 y 5, el 100% coincide con que los ratios financieros incluidos en FILIGRANT facilitan el cálculo de otros indicadores financieros, lo que facilita la realización del análisis clásico para conocer la situación financiera de las empresas. En el caso de la pregunta 6, el 100% de los encuestados considera que es útil vincular la información financiera de las empresas con información procedente de fuentes externas, es este sentido, uno de los encuestados proporcionó su opinión al respecto, tal opinión está incluida en la Tabla 20 y es discutida posteriormente.

Con respecto a las preguntas 7 y 8 del cuestionario, el 100% de los encuestados también presenta coincidencias en sus respuestas, las cuales, indican que la manera en la que se presenta la información financiera a través de FILIGRANT, es decir, de forma agrupada, con indicadores adicionales y complementada con información de otras fuentes de datos, puede facilitar la toma decisiones. Así mismo, fundamentándose en la diversidad que existe en los formatos para la publicación de información financiera, el 100% de los encuestados considera necesario obtener este tipo de información en un formato unificado y conocido por todas las partes, de manera que todos hagan uso de un modelo de publicación de datos financieros común, aunado a este argumento, y con relación a la pregunta 9, el 100% de los encuestados indicaron que la iniciativa FILIGRANT supone una mejora cualitativa en la representación de la información financiera en la Web.

De los 10 expertos que dieron respuesta al cuestionario realizado, 7 proporcionaron sus opiniones acerca de FILIGRANT, estas respuestas enriquecen el trabajo de tesis realizado, tanto sus opiniones como sus recomendaciones, principalmente estas últimas, son consideradas para las futuras líneas de investigación descritas en el capítulo siguiente.

RESPUESTAS A PREGUNTA 10 DEL CUESTIONARIO	
No. Encuestado	**OPINIONES y RECOMENDACIONES**
1	Interesante iniciativa, los ratios por sí solos no son útiles para analizar una empresa, pero como parte de alguna fórmula derivada de ellos sí podrían ser de ayuda para analizar financieramente una empresa.
2	-
3	Pienso que los ratios financieros presentados son útiles para realizar cálculos financieros que ayuden a analizar la situación de las empresas, pero son muchos. Mi recomendación sería unificarlos o agruparlos en conceptos similares.
4	Su propuesta es una buena aportación para el análisis de la situación financiera de las empresas, principalmente la sección que permite calcular indicadores adicionales trimestrales. Le recomendaría añadir más indicadores de este tipo, por ejemplo indicadores de actividad o rentabilidad financiera de las empresas.
5	Considero que su proyecto es útil para buscar ratios financieros publicados por las empresas americanas, con el agregado de mostrar graficas y calcular algunos indicadores adicionales. Algunos ratios financieros son semejantes, le sugiero agruparlos y añadir más indicadores adicionales. Además, sería bueno añadir empresas de otros países.
6	El proyecto es una mejora en la manera de visualizar los datos financieros de forma gráfica y complementado con tablas. Los indicadores adicionales son de utilidad al momento de analizar la información básica de las empresas y obtener información adicional de las empresas, sirve para ahorrar tiempo de búsqueda en Google.
7	-
8	El proyecto proporciona una alternativa para facilitar el análisis financiero de las empresas. Con la aplicación de las fórmulas de contabilidad adecuadas, se podrían tomar algunas decisiones y quizás algunas tendencias de las empresas, gracias a que su trabajo ofrece datos financieros de años anteriores. Sólo falta organizar en grupos los ratios financieros porque hay varios semejantes e incluir empresas de países que no sean Estados Unidos.
9	Muy buena propuesta para almacenar y organizar los datos financieros de las empresas, sólo pediría que los ratios se agrupen de acuerdo a las diferentes categorías para facilitar su manejo.
10	-

Tabla 20. Opiniones y recomendaciones proporcionadas por los expertos

Al analizar las opiniones descritas por los expertos, se encuentran coincidencias referentes a que los ratios incluidos en FILIGRANT, pueden ser aplicados en fórmulas de carácter financiero, tal y como ocurre con los indicadores financieros adicionales incluidos en la propia herramienta, y como se ha demostrado en el cálculo de la Razón de deuda de las empresas Wal-Mart y Costo del caso de estudio (véase Secciones 6.3.4.1 y 6.3.4.2) para facilitar el análisis de las empresas y servir de apoyo a la toma de decisiones tanto manual como automatizada.

Respecto a la funcionalidad que ofrece FILIGRANT, para obtener información adicional, sólo se obtuvo la opinión del encuestado número 6, que enfatizó en el ahorro de tiempo que esta herramienta le ofrece para la búsqueda de información adicional de las

empresas. Adicionalmente, el encuestado 9 menciona que FILIGRANT es una buena propuesta para almacenar y organizar la información financiera de las empresas.

Las opiniones proporcionadas por los encuestados ayudan a confirmar la validación de la hipótesis *H3*, que además de ser validada mediante la estructuración de los datos financieros a través de un formato semántico basado en las taxonomías financieras que integran al modelo semántico Mixto, el 100% de los expertos indicó que la propuesta que se presenta en este trabajo de tesis es una mejora cualitativa en la representación de los datos financieros.

En lo que se refiere a las recomendaciones, la mayoría de los encuestados coinciden con que es necesario agrupar los ratios financieros que son similares, además de añadir más indicadores financieros adicionales. Como ya se ha mencionado, estas recomendaciones son descritas a profundidad en la sección de Futuras líneas de investigación.

6.5 Conclusión del proceso de validación

El presente capítulo se diseñaron y ejecutaron una serie de experimentos que facilitaron la validación de las hipótesis planteadas en este trabajo de tesis, así como el cumplimiento de los objetivos establecidos en la misma.

En el primer experimento se realizó un análisis comparativo entre los modelos semánticos presentados en este trabajo de tesis, es decir, el modelo semántico Mixto y el modelo Entidad-Atributo-Valor. Para llevar a cabo este análisis, primero se poblaron ambos modelos semánticos solventando así la hipótesis ***H1***. Posteriormente, se realizó el diseño y ejecución de una serie de consultas basadas en SPARQL que después de ser procesadas, permitieron obtener los tiempos de recuperación de datos para cada modelo, permitiendo demostrar que el modelo semántico Mixto, es el que mejores tiempos de recuperación de datos obtuvo. Por lo que el resto de las hipótesis, fueron validadas utilizando este modelo.

Después de la elección del modelo semántico Mixto, se prosiguió con los experimentos necesarios para el descubrimiento de datos financieros en la Web. Para la ejecución de estos experimentos, primero se analizaron los marcos de trabajo LIMES y Silk para el descubrimiento de enlaces en la LOD cloud a través de DBpedia. El resultado de este análisis permitió elegir a LIMES como el marco de trabajo adecuado para dar inicio con los experimentos que permitieron el descubrimiento de enlaces en DBpedia. Posteriormente, se delimitó la cantidad de empresas a descubrir en DBpedia, esta delimitación consistió en

establecer un subconjunto de 500 empresas correspondientes al índice S&P500, a partir del número total de empresas almacenadas en el conjunto de datos financieros inspirado en Linked Data.

Una vez delimitado el subconjunto de empresas, mediante LIMES se definieron y ejecutaron los experimentos para el descubrimiento de enlaces en DBpedia con y sin la distancia *Levenshtein* y el uso de *Stopwords,* que permitan obtener enlaces con información relacionada con el subconjunto de las empresas seleccionadas. De los resultados obtenidos, se eligió el que proporcionó la mayor cantidad de enlaces relacionados con el subconjunto de empresas establecido. A este resultado, se le aplicaron las métricas *Precision and Recall,* mediante las que se identificaron los enlaces relevantes, los falsos positivos y los enlaces falsos negativos en DBpedia. Los enlaces falsos positivos, fueron descartados por no aportar información relevante, sin embargo, los enlaces falsos negativos fueron revisados manualmente ya que son enlaces que por algún motivo no fueron descubiertos durante la ejecución de los experimentos, peor que sí contienen información relacionada con las 500 empresas del S&P500. La realización de estos experimentos, permitió comprobar que los datos procesados en la base de conocimientos financieros generada a partir del modelo semántico Mixto, permiten la reutilización de datos con terceros a través de Linked Data, que significa que con estos experimentos se validó la hipótesis ***H2***.

La hipótesis ***H3*** fue comprobada mediante el uso de tecnologías semánticas que permitieron estructurar y transformar en un formato semántico a los datos financieros extraídos a partir de los Estados financieros XBRL para ser publicados en la Web. Esta validación fue corroborada por 10 expertos que coincidieron en que la iniciativa presentada en este trabajo de tesis doctoral, es una mejora cualitativa en la representación de los datos financieros.

Por otra parte, la hipótesis ***H4***, involucró un caso de estudio en el que asesores del banco BBVA Bancomer a través de su departamento de Fondos de inversión, analizan y comparan la situación financiera de las empresas Wal-Mart Stores Inc. y Costco Wholesale Corp., para recomendar a un usuario de nombre Andrés Arroyo Ramírez, la empresa en la que más le conviene realizar una inversión.

El proceso de recomendación, requirió el desarrollo de un visualizador de datos nombrado FILIGRANT. Con el uso de FILIGRANT, los asesores del departamento de Fondos de inversión obtuvieron las gráficas y los datos necesarios para analizar, calcular e

interpretar cada uno de los indicadores financieros siguientes, Razón corriente, Capital de trabajo, Prueba ácida y Razón de deuda. Los resultados obtenidos, facilitaron a los asesores financieros recomendarle a Andrés la empresa Costco para la realización de su inversión. Hasta este momento, con el caso de estudio presentado, se validó la sección de la hipótesis ***H4*** que indica que una base de conocimientos financieros fundamentada en Linked Data, favorece el análisis el análisis fundamental financiero para apoyar la toma de decisiones de forma manual. Con relación a esta hipótesis, también se validó que la base de conocimientos financieros fundamentada en Linked Data, da soporte a la toma de decisiones automatizada. En este sentido, a través de las heurísticas implementadas en FILIGRANT, se compararon los resultados de los indicadores financieros trimestrales obtenidos para el Capital de trabajo y la Prueba ácida de las empresas Wal-Mart y Costco, el resultado, fue la recomendación automática de FILIGRANT para invertir en la empresa Costco.

Finalmente, para confirmar la validación del modelo semántico Mixto inspirado en los principios de Linked Data y por consiguiente, la validación de las hipótesis planteadas en este trabajo de tesis, se formuló un cuestionario publicado en la Web de FILIGRANT. Este cuestionario fue respondido por 10 expertos en finanzas y contabilidad, entre los que se encontraban asesores financieros, economistas, contadores públicos, especialistas en comercio electrónico y administradores de empresas, entre otros. Las respuestas proporcionadas por los expertos, permitieron respaldar el cumplimiento de los objetivos descritos al principio del presente capítulo, y por lo tanto, sustentar los resultados y conclusiones obtenidos durante el proceso de validación del modelo semántico Mixto.

El proceso de validación de los modelos semánticos de datos financieros Inspirados en Linked Data presentados en este trabajo, permite concluir que el modelo semántico Mixto es el más adecuado para la correcta validación de las hipótesis planteadas en esta tesis, adicionalmente, esta investigación proporciona conclusiones más extensas y una serie de posibles líneas de investigación que son descritas más a detalle en el capítulo siguiente.

Capítulo 7

Conclusiones y futuras líneas de investigación

Resumen. Durante la investigación llevada a cabo en este trabajo de tesis doctoral, se realizaron diversas tareas que permitieron cumplir con los objetivos generales y específicos estipulados en el primer capítulo del presente documento. Así mismo, estas tareas facilitaron la validación de las hipótesis de investigación planteadas, lo que sirve de indicio para culminar con este trabajo a través del presente capítulo, en el que incluyen las conclusiones finales, se proporciona un repaso de las principales aportaciones realizadas y se finaliza con la propuesta de un conjunto de futuras líneas de investigación con las que se puede dar continuidad al trabajo presentado.

7. Conclusiones y futuras líneas de investigación

En este capítulo, se presentan las conclusiones finales de la investigación expuesta en la esta tesis doctoral. En estas conclusiones se incluye un repaso de las principales aportaciones realizadas en la misma, posteriormente se propone un conjunto de futuras líneas de investigación con las que se puede dar continuidad al trabajo realizado.

7.1 Conclusiones

El punto de partida con el que dio inicio el presente documento, consistió en involucrar al lector en el dominio de las finanzas, contabilidad, estándares para la publicación de informes financieros, análisis fundamental financiero, integración de datos, apoyo a la toma de decisiones, modelos semánticos y Linked Data (Linked Open Data), que en esencia son elementos clave para la realización del modelo semántico Mixto inspirado en los principios de Linked Data presentado a lo largo de esta tesis doctoral. Entre otros, cada uno de estos temas fue abordado más a detalle en el Estado del arte para dotar al trabajo de una base sólida y fundamentada, en el que también se analizaron y compararon diversos trabajos relacionados con esta iniciativa teniendo como conclusión, que el modelo semántico Mixto que se presenta en este trabajo de tesis favorece la generación de una base de conocimientos financieros ligera, escalable y reutilizable que se inspira en los principios de Linked Data y que es una alternativa útil para complementar iniciativas ya existentes, proporcionándoles un conjunto de características nuevas y destacadas mediante las que se abordan retos reales de la integración de datos financieros y sirve de aporte para ayudar a subsanar la necesidad de construir un ecosistema financiero basado en el uso de los estándares Web actuales.

Para llevar a cabo la investigación expuesta en este documento, se incluyó una serie de procedimientos de carácter técnico-científico que incluyeron la definición de taxonomías financieras, el aislamiento, categorización y clasificación de aquellos elementos económico-financieros provenientes de distintas fuentes de información disponibles en la Web, para transformarlos en datos financieros explorables, leídos e interpretados (computables) por un ordenador con el fin de generar conocimiento financiero, mejorar la calidad estructural de los datos, enlazar los datos con fuentes de información externa y apoyar a la toma de decisiones tanto automatizada como por parte de las personas. Para realizar cada uno de estos procedimientos, fue necesario delimitar la investigación a través de la definición de las hipótesis adecuadas e interesantes que sirvieron de guía principal para la obtención de resultados que pudieran ser relevantes para la comunidad investigadora. El proceso de

validación de las hipótesis planteadas, comprendió seis fases basadas en los ciclos de vida de Linked Data y en las que se incluyó el uso de tecnologías semánticas para extraer y transformar los datos contenidos en los Estados financieros basados en el estándar XBRL publicados en la Web. La conclusión de la aplicación del proceso de validación de las hipótesis, fue la obtención de dos modelos semánticos inspirados en Linked Data, el primero es el modelo Entidad-Atributo-Valor o EAV y el segundo es el modelo semántico Mixto.

A partir de los modelos semánticos obtenidos, se pobló una base de conocimientos financieros inspirada en Linked Data, en la que cada modelo tuvo su propio grafo, esto favoreció la validación de la primer hipótesis planteada en este trabajo de tesis, sin embargo, se requirió identificar el modelo semántico que más adecuado para la validación del resto de las hipótesis. En ese sentido, tras realizar unos experimentos de benchmarking para medir el tiempo de recuperación de datos en ambos modelos, se concluyó que el modelo semántico Mixto es la mejor opción para continuar con los experimentos que permitieran el cumplimiento de los objetivos y la validación de las hipótesis planteadas en este trabajo de tesis. El primero de estos experimentos concluyó con la validación de la segunda hipótesis planteada en este trabajo y consistió en descubrir enlaces relacionados para vincular la información almacenada en el conjunto de datos financieros con información externa, en este experimento se vinculó con DBpedia, sin embargo, el conjunto de datos puede ser vinculado con otras fuentes de información que pudieran ser consideradas como relevantes. Los siguientes experimentos, se centraron en validar la tercera y cuarta hipótesis de este trabajo de tesis, para cumplir con este propósito, se desarrolló la herramienta de apoyo FILIGRANT. Mediante FILIGRANT, se publicó en la Web un cuestionario disponible para expertos en finanzas y contabilidad, las respuestas aportadas por estos expertos permitieron concluir que la iniciativa presentada en este trabajo de tesis, es una mejora cualitativa en la representación de los datos financieros que se publican en la Web. La cuarta hipótesis, también fue validada mediante el uso FILIGRANT, con el que se visualizaron, calcularon, analizaron e interpretaron determinados indicadores financieros aplicados a las empresas Wal-Mart y Costco, lo que permitió realizar una recomendación para invertir en Costco, aunado a esto, mediante las heurísticas de FILIGRANT se procesó automáticamente un par de recomendaciones que indicaron la conveniencia de invertir en Costco, permitiendo concluir que el modelo semántico Mixto facilita la toma de decisiones de manera manual y automatizada.

Todo el proceso de investigación, experimentación, validación y conclusiones obtenidas en cada capítulo de este trabajo de tesis doctoral, permite concluir de manera general que a través del modelo semántico Mixto inspirado en los principios de Linked Data, se proporciona una alternativa de solución a las necesidades o limitaciones actuales concernientes a los Estados financieros basados en XBRL publicados por parte de las empresas, que se han centrado en la manera en la que son presentados los datos financieros a los usuarios, sin aportar funcionalidades adicionales a la estructuración de los datos financieros mediante esquemas y taxonomías con una semántica débil. Expuesto lo anterior, a través del modelo semántico Mixto se proporciona una base de conocimientos financieros reutilizable que proporciona fácil acceso a sus datos mediante el manejo de protocolos asociados a Internet como el HTTP para la navegación e interconexión con otras fuentes de datos, además de proporcionar la capacidad para la realización de procesos de inferencia y el procesamiento de cálculos para el análisis fundamental financiero, cuyos resultados sirven de apoyo en la toma de decisiones tanto manual como automatizada.

7.2 Líneas futuras de investigación

La investigación efectuada ha propiciado una serie de posibles líneas futuras de investigación que permiten ampliar el trabajo realizado a lo largo de todo el proceso de creación de este trabajo de tesis doctoral. Las primeras líneas futuras a considerar son las recomendaciones proporcionadas por los especialistas en finanzas y contabilidad que dieron respuesta al cuestionario publicado en la Web de FILIGRANT. De acuerdo con lo expresado, cuatro de los encuestados recomendaron agrupar los ratios financieros que son similares, esto podría mejorar su organización de mejor manera para facilitar su manejo, adicionalmente, dos encuestados sugirieron añadir más indicadores financieros adicionales, esto podría ampliar las opciones para analizar la situación de financiera de las empresas y ofrecer más alternativas para apoyar en la toma de decisiones. Finalmente, dos de las recomendaciones proporcionadas por los expertos señalan la necesidad de añadir empresas de otros países que no sean de Estados Unidos. Considerando esta recomendación, el sistema EDGAR de la U.S. SEC. es la fuente que más Estados financieros XBRL tiene almacenados lo que significa que hasta el momento, es la fuente de información financiera más viable para la implementación del modelo semántico Mixto inspirado en Linked Data presentado en este trabajo de tesis. No obstante, existen otras organizaciones como la Comisión Nacional del Mercado de Valores (CNMV) en España, que está optando por publicar los estados financieros de las organizaciones empresariales bajo el estándar XBRL,

esto la convertiría en una opción muy viable para la implementación del modelo semántico Mixto.

Dentro de otras posibles líneas futuras de investigación, el conjunto de datos financieros podría vincularse a demás de DBpedia, con otras fuentes externas de información que pudieran ser relevantes, la parte interesante de esta vinculación sería la disminución de la identificación manual de las páginas Web que verdaderamente estén relacionadas con las empresas contenidas en el conjunto de datos financieros. Se trata de una línea de investigación compleja que posiblemente requiera la aplicación de técnicas y algoritmos de *Text Mining* o *Natural Language Processing* (NLP), pero que sin duda, podría facilitar el descubrimiento de enlaces de paginas publicadas en la Web, cuya información esté relacionada con el conjunto de datos financieros presentado. Siguiendo con la aplicación de técnicas y algoritmos de *Text Mining* y NLP, éstas podrían aplicarse para extraer datos financieros de fuentes distintas a XBRL, como Estados financieros publicados en formato PDF, por citar un ejemplo. La transformación de los datos contenidos en este tipo de documentos, aplicando el modelo semántico Mixto, podría ampliar la cantidad de información almacenada en el conjunto de datos financieros.

Otra línea futura de investigación posible, sería mediante la aplicación de algoritmos de *Sentiment Analysis* sobre las opiniones que la gente expresa a través de las redes sociales (social media) acerca de las empresas almacenadas en el conjunto de datos financieros utilizando. Los resultados obtenidos, podrían ser correlacionados con los resultados de ciertos indicadores financieros que permitan verificar si las opiniones de las personas, corresponden o tienen cierta influencia en la situación financiera de las empresas.

Son diversas las líneas futuras de investigación que se pueden derivar a partir del trabajo de tesis presentado en este documento, en esta sección se han presentado las que se consideran, podrían complementar de manera relevante el trabajo de tesis doctoral presentado. En el apartado siguiente, se proporcionan los anexos relacionados con esta iniciativa y es mediante el que se da fin al presente documento.

Anexos

Resumen. En el presente apartado, se proporciona información relacionada con los productos académicos obtenidos y publicados durante la investigación realizada en este trabajo doctoral, así como una lista de acrónimos incluidos a lo largo de las descripciones proporcionadas el mismo.

Publicaciones realizadas

Como consecuencia de esta investigación, se realizaron las siguientes publicaciones en ámbitos científicos relacionados con el enfoque presentado en este trabajo de tesis doctoral:

- Radzimski, M., **Sánchez-Cervantes, J. L.,** Garcia-Crespo, A., & Temiño-Aguirre, I. (2014). **Intelligent Architecture for Comparative Analysis of Public Companies Using Semantics and XBRL Data.** *International Journal of Software Engineering and Knowledge Engineering,* 24(05), 801–823. doi: 10.1142/S0218194014500314
- Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M. Á., Moreno, V., Fraga, A., & **Sánchez-Cervantes, J. L.** (2014). **Feature-based opinion mining through ontologies.** *Expert Systems with Applications*, 41(13), 5995-6008. doi: 10.1016/j.eswa.2014.03.022
- Radzimski, M., **Sánchez-Cervantes, J. L.,** López-Cuadrado, J. L., & García-Crespo, Á. (2014 May). **Predicting Stocks Returns Correlations Based on Unstructured Data sources**. *Second International Workshop on Finance and Economics on the Semantic Web (FEOSW 2014).*
- **Sánchez-Cervantes, J. L.,** Hernández-Chan, G. S., Radzimski, M., Gómez-Berbís, J. M., & García-Crespo, Á. (2013, September 29). **Discovering and Linking Financial Data on the Web.** *In DATA ANALYTICS 2013, The Second International Conference on Data Analytics* (pp. 36-40).
- Colomo-Palacios, R., **Sánchez-Cervantes, J. L.,** Alor-Hernández, G., & Rodríguez-González, A. (2012). **Linked Data: Perspectives for IT Professionals.** *International Journal of Human Capital and Information Technology Professionals (IJHCITP), 3*(3), 1-12. doi:10.4018/jhcitp.2012070101
- Radzimski, M., **Sánchez-Cervantes, J. L.,** Rodríguez-González, A., Gómez-Berbís, J. M., & García-Crespo, Á. (2012). **FLORA–Publishing Unstructured Financial Information in the Linked Open Data Cloud.** *In International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)* (pp. 27-28).

Acrónimos

ADR's	American Depositary Receipts
AI	Artificial Intelligence
AICPA	American Institute of Certified Public Accountants
AIFB	Institute of Applied Informatics and Formal Description Methods
API	Application Programming Interface
ASN.1	Abstract Syntax Notation One
CDM	Common Data Model
CLIPS	C-Language Integrated Production System
CNMV	Comisión Nacional del Mercado de Valores
CSS	Cascading Style sheets
CWM	Closed World Machine
DARPA	Agent Markup Language + OIL
DAWG	W3C RDF Data Access Working Group
DBMS	DataBase Management System
DL	Description Logic
DOM	Document Object Model
DTD	Document Type Definitions
E-R	Model, Entity–Relationship Model
EAV	Entity–Attribute–Value model
EDGAR System	Electronic Data Gathering Analysis and Retrieval System
F-Logic	Frame Logic
FASB	Financial Accounting Standards Board
FTP	File Transfer Protocol
GAAP	Generally Accepted Accounting Principles
GNU	General Public License
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IAS	International Accounting Standards
IASB	International Accounting Standards Board
IASC	International Accounting Standards Committee
IFRS	International Financial Reporting Standards
IFRS	International Financial Reporting Standards
IOSCO	Organization of Securities Commission
ISC	International Organization for Standardization
ISI	Information Sciences Institute
KIF	Knowledge Interchange Format
KMI	Knowledge Media Institute
KR Paradigm	Knowledge Representation paradigm

KSL	Knowledge Systems Laboratory
LIMES	LInk discovery framework for MEtric Spaces
LISP	LISt Processing Language
LOD cloud	Linking Open Data cloud
N3	Notation 3
NIC	Normas Internacionales de Contabilidad
NIIF	Normas Internacionales de Información Financiera
OCML	Operational Conceptual Modelling Language
OICV	Organización Internacional de Comisiones de Valores
OIL	Ontology Interchange Language
OKBC	Open Knowledge Based Connectivity
OML	Ontology Markup Language
OntoSaurus	OntoSaurus Ontology Tool
OODM	Object-Oriented Data Model
ORDBMS	Object Relational Database Management System
OWL	Web Ontology Language
OWL 2	Web Ontology Language 2
PCGA de EE.UU	Principios de Contabilidad Generalmente Aceptados en Estados Unidos
RDBMS	Relational Database Management System
RDF	Resource Description Framework
RDF(S)	Resource Description Framework Schema
RDQL	RDF Data Query Language
RDQL	RDF Data Query Language
REST	Representational State Transfer
RQL	Resource Query Language
S&P500	Standard & Poor's 500
SGML	Standard Generalized Markup Language
SHOE	Simple HTML Ontology Extensions
SOAP	Simple Object Access Protocol
SPARQL	Protocol and RDF Query Language
SQL	Structured Query Language
U.S. SEC	USA-Securities and Exchange Commission
UML	Unified Modeling Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
US-GAAP	US-Generally Accepted Accounting Principles
W3C	World Wide Web Consortium
WAB	WebODE, Axiom Builder
XBRL	eXtensible Business Reporting Language
XFRML	eXtensible Financial Reporting Markup Language

XML	Extensible Markup Language
XOL	XML-Based Ontology Exchange Language
XSD	XML Schema Definition
XSLT	eXtensible Stylesheet Language Transformations

Bibliografía

Aasman, J. (2006). *Allegro Graph: RDF Triple Database*.

Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the Web: from relations to semistructured data and XML* (pp. 32–33). Morgan Kaufmann.

Abrial, J.-R. (1974). Data Semantics. In *IFIP Working Conference Data Base Management* (pp. 1–60).

Acedo-Peñalva, F. (2006). Principales diferencias entre las NIIF y las US GAAP. *Harvard Deusto Finanzas Y Contabilidad*, (70), 46–55.

Acosta-González, E., & Fernández-Rodríguez, F. (2014). Forecasting Financial Failure of Firms via Genetic Algorithms. *Computational Economics*, *43*(2), 133–157. doi:10.1007/s10614-013-9392-9

Ahmadpour, A., & Bodaghi, A. (2012). The effects of XBRL on financial transparency. *International Journal of Information Science and Management (IJISM)*, 65–76.

Albert, R., Jeong, H., & Barabasi, A.-L. (1999). Internet: Diameter of the World-Wide Web. *Nature*, *401*(6749), 130–131. Retrieved from http://dx.doi.org/10.1038/43601

Alexander, N., & Ravada, S. (2006). RDF Object Type and Reification in the Database. In *Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on* (p. 93). doi:10.1109/ICDE.2006.126

Allemang, D. (2010). Semantic Web and the Linked Data Enterprise. In D. Wood (Ed.), *Linking Enterprise Data SE - 1* (pp. 3–23). Springer US. doi:10.1007/978-1-4419-7665-9_1

Allen, M. F., & Cote, J. (2005). Creditors' Use of Operating Cash Flows: An Experimental Study. *Journal of Managerial Issues*, *17*(2), 98–211.

Alvarez, J. M., Labra, J. E., Cifuentes, F., Alor-hérnandez, G., Sánchez, C., & Luna, J. A. G. (2012). Towards a pan-european e-procurement platform to aggregate, publish and search public procurement notices powered by linked open data: the Moldeas approach. *International Journal of Software Engineering and Knowledge Engineering*, *22*(03), 365–383. doi:10.1142/S0218194012400086

Anhøj, J. (2003). Generic design of Web-based clinical databases. *Journal of Medical Internet Research*, *5*(4), e27. doi:10.2196/jmir.5.4.e27

Antoniou, G., Franconi, E., & Van-Harmelen, F. (2005). Introduction to semantic web ontology languages. In *Reasoning web* (pp. 1–21). Springer. doi:10.1007/11526988_1

Antoniou, G., & Van-Harmelen, F. (2004). Web ontology language: Owl. In *Handbook on ontologies* (pp. 67–92). Springer. doi:10.1007/978-3-540-24750-0_4

Arndt, H.-K., Isenmann, R., Brosowski, J., Thiessen, I., & Marx Gómez, J. (2006). Sustainability reporting using the extensible business reporting language (XBRL). *Tochtermann & A. Scharl (Eds.), Managing Environmental Knowledge*, 75–82.

Arpírez, J. C., Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2001). WebODE: a scalable workbench for ontological engineering. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 6–13). doi:http://dx.doi.org/10.1145/500737.500743

Arpírez, J. C., Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). WebODE in a nutshell. *AI Magazine*, *24*(3), 37. doi:http://dx.doi.org/10.1609/aimag.v24i3.1717

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, … P. Cudré-Mauroux (Eds.), *The Semantic Web SE - 52* (Vol. 4825, pp. 722–735). Springer Berlin Heidelberg. doi:10.1007/978-3-540-76298-0_52

Auer, S., Dietzold, S., Lehmann, J., Hellmann, S., & Aumueller, D. (2009). Triplify: light-weight linked data publication from relational databases. In *Proceedings of the 18th international conference on World wide web* (pp. 621–630). New York, NY, USA: ACM. doi:10.1145/1526709.1526793

Auer, S., Lehmann, J., Ngonga Ngomo, A.-C., & Zaveri, A. (2013). Introduction to Linked Data and Its Lifecycle on the Web. In S. Rudolph, G. Gottlob, I. Horrocks, & F. van Harmelen (Eds.), *Reasoning Web. Semantic Technologies for Intelligent Data Access SE - 1* (Vol. 8067, pp. 1–90). Springer Berlin Heidelberg. doi:10.1007/978-3-642-39784-4_1

Baader, F., Brandt, S., & Lutz, C. (2005). Pushing the EL envelope. In *IJCAI* (Vol. 5, pp. 364–369).

Bachman, C. W., Batchelor, R. E., Beriss, I. M., Blose, C. R., Burakreis, T. I., Valle, V. D., … Werner, G. T. (1969). *Data Base Task Group Report to the CODASYL Programming Language Committee, October 1969*. New York, NY, USA: ACM.

Balaban, M. (1995). The F-logic approach for description languages. *Annals of Mathematics and Artificial Intelligence*, *15*(1), 19–60.

Bartley, J., Chen, A. Y. S., & Taylor, E. Z. (2011). A Comparison of XBRL Filings to Corporate 10-Ks—Evidence from the Voluntary Filing Program. *Accounting Horizons*, *25*(2), 227–245. doi:http://dx.doi.org/10.2308/acch-10028

Bechhofer, S. (2009). OWL: Web ontology language. In *Encyclopedia of Database Systems* (pp. 2008–2009). Springer. doi:10.1007/978-0-387-39940-9_1073

Bechhofer, S., Horrocks, I., Goble, C., & Stevens, R. (2001). OilEd: a reason-able ontology editor for the semantic web. In *KI 2001: Advances in Artificial Intelligence* (pp. 396–408). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/3-540-45422-5_28

Beckett, D. (2002). The design and implementation of the Redland RDF application framework. *Computer Networks*, *39*(5), 577–588. doi:10.1016/S1389-1286(02)00221-9

Beckett, D. (2014). RDF 1.1 N-Triples. *W3C Recommendation*. Retrieved June 10, 2014, from http://www.w3.org/TR/2014/REC-n-triples-20140225/

Beckett, D., & Barstow, A. (2001). N-Triples. *W3C RDF Core WG Internal Working Draft*. Retrieved June 10, 2014, from http://www.w3.org/2001/sw/RDFCore/ntriples/

Beckett, D., & Berners-Lee, T. (2011). Turtle-Terse RDF Triple Language. *W3C Team Submission*, *14*. Retrieved from http://www.w3.org/TR/2011/WD-turtle-20110809/

Berendt, B., Hotho, A., Mladenic, D., Van Someren, M., Spiliopoulou, M., & Stumme, G. (2004). *A roadmap for web mining: From web to semantic web* (pp. 1–22). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/978-3-540-30123-3_1

Berners-Lee, T. (2009). Linked Data - Design Issues. *Linked Data*. Retrieved October 08, 2013, from http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2013). Relational Databases on the Semantic Web. Retrieved from http://www.w3.org/DesignIssues/RDB-RDF.html

Berners-Lee, T., & Connolly, D. (1998). Notation3 (N3): A readable RDF syntax. *Notation3 (N3): A readable RDF syntax*. Retrieved June 10, 2014, from http://www.w3.org/TeamSubmission/n3/

Berners-Lee, T., & Connolly, D. (2000). CWM-closed world machine. *Internet: Http://www. w3. org/2000/10/swap/doc/cwm. Html*. Retrieved from http://www.w3.org/2000/10/swap/doc/cwm.html

Berners-Lee, T., & Connolly, D. (2002). Euler proof mechanism. *Euler Proof Mechanism*. Retrieved April 03, 2014, from http://www.agfa.com/w3c/euler

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 28–37.

Berners-Lee, T. J. (1992). The world-wide web. *Computer Networks and ISDN Systems*, *25*(4-5), 454–459. doi:10.1016/0169-7552(92)90039-S

Berrueta, D., & Phipps, J. (2008). Best Practice Recipes for Publishing RDF Vocabularies-W3C Working Group Note. *W3C Working Group Note*. Retrieved April 12, 2014, from http://www.w3.org/TR/swbp-vocab-pub/

Binstock, C., Hoffman, C., Egmond, R., & Walenga, W. (2005). *Comparing XML and XBRL*. Retrieved from http://www.ubmatrix.com/Documents/XBRLComparedToXML-2005-07-06 (4).pdf

Biron, P., Malhotra, A., & W3C-Group. (2012). XML schema part 2: Datatypes. *World Wide Web Consortium Recommendation REC-xmlschema-2-20041028*. Retrieved October 09, 2013, from http://www.w3.org/TR/xmlschema11-2/

Bizer, C. (2003). D2r map-a database to rdf mapping language. Retrieved from http://wwwconference.org/www2003/cdrom/papers/poster/p004/p4-bizer.html

Bizer, C., & Cyganiak, R. (2006). D2r server-publishing relational databases on the semantic web. In *Poster at the 5th International Semantic Web Conference*.

Bizer, C., Heath, T., Ayers, D., & Raimond, Y. (2007). Interlinking open data on the web. In *Demonstrations Track, 4th European Semantic Web Conference, Innsbruck, Austria*.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi:10.4018/jswis.2009081901

Bizer, C., Jentzsch, A., & Cyganiak, R. (2011). State of the LOD Cloud. *Version 0.3 (September 2011)*, *1803*. Retrieved from http://lod-cloud.net/state/

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(3), 154–165. doi:10.1016/j.websem.2009.07.002

Bizer, C., & Schultz, A. (2009). The Berlin SPARQL Benchmark. *International Journal on Semantic Web and Information Systems*, *5*(2), 1–24. doi:10.4018/jswis.2009040101

Bizer, Cyganiak, R., & Heath, T. (2007). How to publish linked data on the web. Retrieved April 12, 2014, from http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/

Bliss, J. H. (1923). *Financial and operating ratios in management* (pp. 34–38). The Ronald press company.

Bobrow, D. G., & Murphy, D. L. (1967). Structure of a LISP System Using Two-level Storage. *Commun. ACM*, *10*(3), 155–159. doi:10.1145/363162.363185

Bojars, U., Passant, A., Cyganiak, R., & Breslin, J. (2008). Weaving sioc into the web of linked data. In *Linked Data on the Web (LDOW 2008) workshop, in conjunction with WWW 2008 conference*.

Bonner, A. J., & Kifer, M. (1994). An overview of transaction logic. *Theoretical Computer Science*, *133*(2), 205–265. doi:10.1016/0304-3975(94)90190-2

Borah, J. (2002). Conceptual Modeling--The Missing Link of Simulation Development. In *Proceedings of the 2002 Spring Simulation Conference*.

Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente.

Bovee, M., Kogan, A., Nelson, K., Srivastava, R. P., & Vasarhelyi, M. A. (2005). Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL). *Journal of Information Systems*, *19*(1), 19–41. doi:10.2308/jis.2005.19.1.19

Brachman, R. J. (1978). On the epistemological status of semantic networks. *NASA STI/Recon Technical Report N*, *78*, 31337.

Brainard, W. C., & Tobin, J. (1968). Pitfalls in financial model building. *The American Economic Review*, *58*(2), 99–122.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (1998). Extensible markup language (XML). *World Wide Web Consortium Recommendation REC-Xml-19980210. Http://www. w3. org/TR/1998/REC-Xml-19980210*.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brenner, T., & Werker, C. (2007). A taxonomy of inference in simulation models. *Computational Economics*, *30*(3), 227–244. doi:10.1007/s10614-007-9102-6

Brewster, C., & O'Hara, K. (2007). Knowledge representation with ontologies: Present challenges—Future possibilities. *International Journal of Human-Computer Studies*, *65*(7), 563–568. doi:http://dx.doi.org/10.1016/j.ijhcs.2007.04.003

Brickley, D., & Guha, R. V. (2014). RDF Schema 1.1 - W3C Recommendation. Retrieved March 05, 2014, from http://www.w3.org/TR/rdf-schema/

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1-7), 107–117. doi:10.1016/S0169-7552(98)00110-X

Brodie, M. (1984). On the Development of Data Models. In M. Brodie, J. Mylopoulos, & J. Schmidt (Eds.), *On Conceptual Modelling SE - 2* (pp. 19–47). Springer New York. doi:10.1007/978-1-4612-5196-5_2

Broekstra, J., & Kampman, A. (2003). SeRQL: a second generation RDF query language. In *Proc. SWAD-Europe Workshop on Semantic Web Storage and Retrieval* (pp. 13–14).

Broekstra, J., & Kampman, A. (2004). Serql: An rdf query and transformation language. In *Submitted to the International Semantic Web Conference, ISWC* (Vol. 2004).

Broekstra, J., Kampman, A., & Van Harmelen, F. (2002). Sesame: A generic architecture for storing and querying rdf and rdf schema. In *The Semantic Web—ISWC 2002* (pp. 54–68). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/3-540-48005-6_7

Broekstra, J., Klein, M., Decker, S., Fensel, D., van Harmelen, F., & Horrocks, I. (2002). Enabling knowledge representation on the Web by extending RDF Schema. *Computer Networks*, *39*(5), 609–634. doi:10.1016/S1389-1286(02)00217-7

Callao, S., Jarne, J. I., & Laínez, J. A. (2007). Adoption of {IFRS} in Spain: Effect on the comparability and relevance of financial reporting. *Journal of International Accounting, Auditing and Taxation*, *16*(2), 148–178. doi:http://dx.doi.org/10.1016/j.intaccaudtax.2007.06.002

Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., & Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, *39*(3), 385–429. doi:10.1007/s10817-007-9078-x

Carroll, G., & Klyne, J. J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. *W3C Recommendation*. Retrieved October 09, 2013, from http://www.w3.org/TR/rdf-concepts/

Carroll, J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., & Wilkinson, K. (2004). Jena: Implementing the Semantic Web Recommendations. In *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers &Amp; Posters* (pp. 74–83). New York, NY, USA: ACM. doi:10.1145/1013367.1013381

Carterette, B. (2009). Precision and Recall. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems SE - 5050* (pp. 2126–2127). Springer US. doi:10.1007/978-0-387-39940-9_5050

Castellanos-Nieves, D., Fernández-Breis, J. T., Valencia-García, R., Martínez-Béjar, R., & Iniesta-Moreno, M. (2011). Semantic Web Technologies for supporting learning assessment. *Information Sciences*, *181*(9), 1517–1537. doi:http://dx.doi.org/10.1016/j.ins.2011.01.010

Chagas, F., De-Carvalho, C. L., & Da-Silva, J. C. (2008). Semantic Web Support Applications. In *Proceedings of the 2008 Euro American Conference on Telematics and Information Systems* (pp. 27:1–27:4). New York, NY, USA: ACM. doi:10.1145/1621087.1621114

Chaudhri, V. K., Farquhar, A., Fikes, R., Karp, P. D., & Rice, J. P. (1998). Open Knowledge Base Connectivity 2.0. 3 (Proposed). *Artificial Intelligence Center of SRI International and Knowledge Systems Laboratory of Stanford University*.

Chen, P. P.-S. (1976). The Entity-relationship Model&Mdash;Toward a Unified View of Data. *ACM Trans. Database Syst.*, *1*(1), 9–36. doi:10.1145/320434.320440

Cheng, H., Lu, Y.-C., & Sheu, C. (2009). An ontology-based business intelligence application in a financial knowledge management system. *Expert Systems with Applications*, *36*(2), 3614–3622. doi:10.1016/j.eswa.2008.02.047

Christodoulakis, S., Ho, F., & Theodoridou, M. (1986). The Multimedia Object Presentation Manager of MINOS: A Symmetric Approach. *SIGMOD Rec.*, *15*(2), 295–310. doi:10.1145/16856.16884

Clark, K. G. (2005). RDF Data Access Use Cases and Requirements. *Working draft, W3C*. Retrieved January 20, 2014, from http://www.w3.org/TR/rdf-dawg-uc/

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, *13*(6), 377–387. doi:10.1145/362384.362685

Coetzee, P., Heath, T., & Motta, E. (2008). SparqPlug: Generating Linked Data from Legacy HTML, SPARQL and the DOM. In *LDOW*.

Colomo-Palacios, R., Sánchez-Cervantes, J. L., Alor-Hernandez, G., & Rodríguez-González, A. (2012). Linked Data. *International Journal of Human Capital and Information Technology Professionals*, *3*(3), 1–12. doi:10.4018/jhcitp.2012070101

Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, *46*(1), 41–64. doi:10.1016/S0169-023X(02)00195-7

Creamer, G., & Freund, Y. (2010). Using Boosting for Financial Analysis and Performance Prediction: Application to S&P 500 Companies, Latin American ADRs and Banks. *Computational Economics*, *36*(2), 133–151. doi:10.1007/s10614-010-9205-3

Curry, E., Freitas, A., & O'Riáin, S. (2010). The Role of Community-Driven Data Curation for Enterprises. In D. Wood (Ed.), *Linking Enterprise Data SE - 2* (pp. 25–47). Springer US. doi:10.1007/978-1-4419-7665-9_2

Curry, E., O'Donnell, J., Corry, E., Hasan, S., Keane, M., & O'Riain, S. (2013). Linking building data in the cloud: Integrating cross-domain building data using linked data. *Advanced Engineering Informatics*, *27*(2), 206–219. doi:10.1016/j.aei.2012.10.003

Cyganiak, R., & Bizer, C. (2008). Pubby-A Linked Data Frontend for SPARQL Endpoints. Retrieved April 14, 2014, from http://wifo5-03.informatik.uni-mannheim.de/pubby/

Cyganiak, R., Wood, D., & Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation*. Retrieved June 09, 2014, from http://www.w3.org/TR/rdf11-concepts/

Date, C. J., & Faudón, S. L. M. R. (2001). *Introducción a los sistemas de bases de datos* (7th ed., pp. 13–15). Pearson Publications Company.

Davies, J., Fensel, D., & Harmelen, F. Van. (2003). *Towards the semantic web*. Wiley Online Library.

Debreceny, R., & Gray, G. (2001). The production and use of semantically rich accounting reports on the Internet: XML and XBRL. *International Journal of Accounting Information …*. Retrieved from http://www.sciencedirect.com/science/article/pii/S1467089500000129

Debreceny, R., Gray, G., & Barry, T. (1998). Accounting Information in a Networked World-Resource Discovery, Processing and Analysis. In *American Accounting Association Annual Meeting, at New Orleans*.

Decker, S. (2002). Logic Databases on the Semantic Web: Challenges and Opportunities. In P. Stuckey (Ed.), *Logic Programming SE - 2* (Vol. 2401, pp. 20–21). Springer Berlin Heidelberg. doi:10.1007/3-540-45619-8_2

Decker, S., Erdmann, M., Fensel, D., & Studer, R. (1999). *Ontobroker: Ontology based access to distributed and semi-structured information* (pp. 351–369). Springer US. doi:http://dx.doi.org/10.1007/978-0-387-35561-0_20

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decision tree approach. *Expert Systems with Applications*, *40*(10), 3970–3983. doi:http://dx.doi.org/10.1016/j.eswa.2013.01.012

Dell, M., & Dell, S. (2010). *Canonical Data Model Design Guidelines* (pp. 1–16). Retrieved from http://canonical.ciss.com.br/webcontent/material/CDM_Design_Guidelines.pdf

Der-Meulen, S. Van, Gaeremynck, A., & Willekens, M. (2007). Attribute differences between U.S. GAAP and IFRS earnings: An exploratory study. *The International Journal of Accounting*, *42*(2), 123–142. doi:http://dx.doi.org/10.1016/j.intacc.2007.04.001

DeRose, S., Maler, E., Orchard, D., & Trafford, B. (2000). XML linking language (XLink). *Working Draft WD-xlink-20000221, World Wide Web Consortium (W3C)*. Retrieved October 09, 2013, from http://www.w3pdf.com/W3cSpec/XLink/2/REC-xlink11-20100506.pdf

Dietrich, J., Jones, N., & Wright, J. (2008). Using social networking and semantic web technology in software engineering – Use cases, patterns, and a case study. *Journal of Systems and Software*, *81*(12), 2183–2193. doi:http://dx.doi.org/10.1016/j.jss.2008.03.060

Dietz, J. L. G., & Habing, N. (2004). A meta Ontology for Organizations. In *On the move to meaningful internet systems 2004: OTM 2004 workshops* (pp. 533–543).

Dimitrov, M. (2012). Semantic Technologies and Triplestores for Business Intelligence. In *Business Intelligence* (pp. 139–155). Springer. doi:10.1007/978-3-642-27358-2_7

Dodge, R. (1991). Earnings per share. In *The Concise Guide to Accounting Standards SE - 3* (pp. 13–18). Springer US. doi:10.1007/978-1-4899-7096-1_3

Domingue, J. (1998). Tadzebao and WebOnto: Discussing, browsing, and editing ontologies on the web. In *Eleventh Workshop on Knowledge Acquisition,Modeling and Management*. Banff, Alberta, Canada.

Doolin, B., & Troshani, I. (2004). XBRL: a research note. *Qualitative Research in Accounting & Management*, *1*(2), 93–104. doi:10.1108/11766090410813373

Du, J., & Zhou, L. (2012). Improving financial data quality using ontologies. *Decision Support Systems*, *54*(1), 76–86. doi:10.1016/j.dss.2012.04.016

DuCharme, B. (2011). *Learning Sparql*. O'Reilly Media.

Eco, U. (2001). *C{ó}mo se hace una tesis: T{é}cnicas y procedimientos de investigaci{ó}n, estudio y escritura*. Gedisa.

Efrim-Boritz, J., & No, W. G. (2005). Security in XML-based financial reporting services on the Internet. *Journal of Accounting and Public Policy*, *24*(1), 11–35. doi:10.1016/j.jaccpubpol.2004.12.002

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. (S. E. S. reprint of the original 1st ed. 1993, Ed.) (Vol. 57). CRC press.

Engel, P., Hamscher, W., Advantage, S., Shuetrim, G., vun Kannon, D., & Pryde, C. (2008). Extensible Business Reporting Language (XBRL) 2.1. *Jul*, *2*, 1–165. Retrieved from http://www.xbrl.org/Specification/XBRL-RECOMMENDATION-2003-12-31+Corrected-Errata-2004-04-29.pdf

Erling, O., & Mikhailov, I. (2009). RDF Support in the Virtuoso DBMS. In *Networked Knowledge-Networked Media* (pp. 7–24). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/978-3-642-02184-8_2

Evans, M. (2005). Should bond markets be more transparent. *International Financial Law Review*, *24*(10), 6.

Farquhar, A., Fikes, R., & Rice, J. (1997). The Ontolingua Server: a tool for collaborative ontology construction. *International Journal of Human-Computer Studies*, *46*(6), 707–727. doi:10.1006/ijhc.1996.0121

FASB. (2002). The Norwalk Agreement. *Memorandum of Understanding – FASB and IASB*, pp. 1–2. Connecticut, USA. Retrieved from http://www.fasb.org/news/memorandum.pdf

Fensel, D., Decker, S., Erdmann, M., & Studer, R. (1998). Ontobroker: the very high idea. In *FLAIRS Conference* (pp. 131–135).

Fensel, D., Horrocks, I., Van Harmelen, F., Decker, S., Erdmann, M., & Klein, M. (2000). OIL in a nutshell. In *Knowledge Engineering and Knowledge Management Methods, Models, and Tools* (pp. 1–16). Springer. doi:10.1007/3-540-39967-4_1

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., & Berners-Lee, T. (1999). Hypertext transfer protocol--HTTP/1.1. RFC 2616, June. Retrieved from http://www.hjp.at/doc/rfc/rfc2616.html

Fields, T. D., Lys, T. Z., & Vincent, L. (2001). Empirical research on accounting choice. *Journal of Accounting and Economics*, *31*(1-3), 255–307. doi:10.1016/S0165-4101(01)00028-3

Freitas, A., Oliveira, J. G., O'Riain, S., da Silva, J. C. P., & Curry, E. (2013). Querying linked data graphs using semantic relatedness: A vocabulary independent approach. *Data & Knowledge Engineering*, *88*, 126–141. doi:10.1016/j.datak.2013.08.003

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. doi:http://dx.doi.org/10.1006/jcss.1997.1504

Gangemi, A. (2005). Ontology design patterns for semantic web content. In *The Semantic Web--ISWC 2005* (pp. 262–276). Springer. doi:http://dx.doi.org/10.1007/11574620_21

García, R., & Gil, R. (2010a). Linking XBRL Financial Data. In D. Wood (Ed.), *Linking Enterprise Data SE - 6* (pp. 103–125). Springer US. doi:10.1007/978-1-4419-7665-9_6

García, R., & Gil, R. (2010b). Triplificating and linking XBRL financial data. In *Proceedings of the 6th International Conference on Semantic Systems* (pp. 3:1–3:8). New York, NY, USA: ACM. doi:10.1145/1839707.1839711

Gardner, D., Knuth, K. H., Abato, M., Erde, S. M., White, T., DeBellis, R., & Gardner, E. P. (2001). Common Data Model for Neuroscience Data and Data Model Exchange. *Journal of the American Medical Informatics Association*, *8*(1), 17–33. doi:10.1136/jamia.2001.0080017

Genesereth, M. R., & Fikes, R. E. (1992). Knowledge interchange format-version 3.0: Reference manual.

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., … Tu, S. W. (2003). The evolution of Protégé: an environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, *58*(1), 89–123. doi:10.1016/S1071-5819(02)00127-1

Georgeff, M. P., Lansky, A. L., & Bessiere, P. (1985). A Procedural Logic. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 516–523). Retrieved from http://www.ijcai.org/Past Proceedings/IJCAI-85-VOL1/PDF/099.pdf

Gerdes, J. (2003). EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database. *Decision Support Systems*, *35*(1), 7–29. doi:10.1016/S0167-9236(02)00096-9

Ginsberg, M. L. (1991). Knowledge interchange format: The KIF of death. *AI Magazine*, *12*(3), 57. doi:http://dx.doi.org/10.1609/aimag.v12i3.903

Goldfarb, C. F. (1990). *The SGML Handbook*. (Y. Rubinsky, Ed.). Oxford University Press.

Gomaa, M. I., Markelevich, A., & Shaw, L. (2011). Introducing {XBRL} through a financial statement analysis project. *Journal of Accounting Education*, *29*(2–3), 153–173. doi:http://dx.doi.org/10.1016/j.jaccedu.2011.12.001

Gómez-Berbís, J., García-Sánchez, F., Valencia-García, R., Toma, I., & Moreno, C. (2009). SONAR: A Semantically Empowered Financial Search Engine. In J. Mira, J. Ferrández, J. Álvarez, F. Paz, & F. J. Toledo (Eds.), *Methods and Models in Artificial and Natural Computation. A Homage to Professor Mira's Scientific Legacy SE - 42* (Vol. 5601, pp. 405–414). Springer Berlin Heidelberg. doi:10.1007/978-3-642-02264-7_42

Gómez-Pérez, A., & Corcho, O. (2002). Ontology languages for the Semantic Web. *Intelligent Systems, IEEE*, *17*(1), 54–60. doi:10.1109/5254.988453

Gonzalez, A. J., & Dankel, D. D. (1993). *The engineering of knowledge-based systems: theory and practice*. (E. Cliffs, Ed.). New Jersey, USA: Prentice hall Englewood Cliffs (NJ).

González, A., & Padilla, J. (1999). Un Esquema Conceptual para Analizar la Validez en las Investigaciones Mediante Encuesta. Metodologia de Encuestas. *Revista de La Sociedad Internacional de Profesionales de La Investigación En Encuestas*, *1*(1), 85–98.

Gordon, E., Jorgensen, B., & Linthicum, C. (2008). *Could IFRS replace US GAAP? A comparison of earnings attributes and informativeness in the US market*.

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., & Sattler, U. (2008). OWL 2: The next step for OWL. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(4), 309–322. doi:10.1016/j.websem.2008.05.001

Gray, G. L., & Miller, D. W. (2009). XBRL: Solving real-world problems. *International Journal of Disclosure and Governance*, *6*(3), 207–223. doi:http://dx.doi.org/10.1057/jdg.2009.8

Grosof, B. (2009). Opportunities for Semantic Web knowledge representation to help XBRL. In *Workshop on Improving Access to Financial Data on the Web. XBRL International and World Wide Web Consortium (W3C)*.

Gruber, T. R. (1992). *Ontolingua: A mechanism to support portable ontologies*. Stanford University, Knowledge Systems Laboratory.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. doi:http://dx.doi.org/10.1006/knac.1993.1008

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5–6), 907–928. doi:http://dx.doi.org/10.1006/ijhc.1995.1081

Haase, P., Broekstra, J., Eberhart, A., & Volz, R. (2004). A comparison of RDF query languages. In *The Semantic Web--ISWC 2004* (pp. 502–517). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/978-3-540-30475-3_35

Hammer, M., & Mc Leod, D. (1981). Database Description with SDM: A Semantic Database Model. *ACM Trans. Database Syst.*, *6*(3), 351–386. doi:10.1145/319587.319588

Harold, E. R. (2004). *XML 1.1 Bible* (3rd ed.). New York, NY, USA: John Wiley & Sons.

Harris, S., Lamb, N., & Shadbolt, N. (2009). 4store: The design and implementation of a clustered RDF store. In *5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2009)* (pp. 94–109).

Harris, T., & Morsfield, S. (2012). An Evaluation of the Current State and Future of XBRL and Interactive Data for Investors and Analysts. Retrieved from http://academiccommons.columbia.edu/catalog/ac:161038

Hartig, O., Bizer, C., & Freytag, J.-C. (2009). Executing SPARQL Queries over the Web of Linked Data. In A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, & K. Thirunarayan (Eds.), *The Semantic Web - ISWC 2009 SE - 19* (Vol. 5823, pp. 293–309). Springer Berlin Heidelberg. doi:10.1007/978-3-642-04930-9_19

Haslhofer, B., & Schandl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. In *International Workshop on Linked Data on the Web*

(LDOW2008), co-located with WWW 2008. Beijing. Retrieved from http://eprints.cs.univie.ac.at/284/

Hassanzadeh, O., Kementsietsidis, A., Lim, L., Miller, R. J., & Wang, M. (2009). LinkedCT: A Linked Data Space for Clinical Trials. Retrieved from http://arxiv.org/abs/0908.0567

Hausenblas, M. (2009). Exploiting Linked Data to Build Web Applications. *IEEE Internet Computing*, *13*(4), 68–73. doi:http://doi.ieeecomputersociety.org/10.1109/MIC.2009.79

Hausenblas, M., Villazón-Terrazas, B., & Hyland, B. (2011). GLD life cycle. *W3C government linked data group, W3C*. Retrieved July 21, 2014, from http://www.w3.org/2011/gld/wiki/GLD_Life_cycle

Hayes, P., Eskridge, T. C., Mehrotra, M., Bobrovnikoff, D., Reichherzer, T., & Saavedra, R. (2005). COE: Tools for collaborative ontology development and reuse. In *Knowledge Capture Conference (KCAP)*.

Heath, T. (2010). Linked data-connect distributed data across the web. Online. Retrieved May 05, 2014, from http://linkeddata.org/

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *1*(1), 1–136.

Heflin, J., Hendler, J., & Luke, S. (1999). SHOE: A knowledge representation language for internet applications. doi:http://hdl.handle.net/1903/1044

Hirst, D. E., & Hopkins, P. E. (1998). Comprehensive Income Disclosures and Analysts' Valuation Judgments. *Journal of Accounting Research*, *36*, 47–75. Retrieved from http://ssrn.com/abstract=145668

HJ, L., & G, B. (1994). UNderstanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *JAMA*, *271*(14), 1103–1108. doi:10.1001/jama.1994.03510380059038

Hobbs, J. R., & Pan, F. (2004). An Ontology of Time for the Semantic Web, *3*(1), 66–85. doi:10.1145/1017068.1017073

Hobbs, J. R., & Pan, F. (2006). Time Ontology in OWL. W3C Working Draft, 27 September 2006. *World Wide Web Consortium*. Retrieved August 25, 2014, from http://www.w3.org/TR/owl-time/

Hoffman, C., & Van-Egmond, R. (2012). *Digital Financial Reporting Using an XBRL-based Model* (pp. 44–73). Retrieved from http://www.xbrl.org/sites/xbrl.org/files/resources/digital_financial_reporting-using_an_xbrl_model.pdf

Horrocks, I., Van-Harmelen, F., & Patel-Schneider, P. F. (2001). *Reference description of the daml+ oil (march 2001) ontology markup language*. Retrieved from http://www.daml.org/2000/12/reference.html

Hosmer, D. W., & Lemeshow, S. (1989). Introduction to the Logistic Regression Model. *Applied Logistic Regression, Second Edition*, 1–30.

Hull, R., & King, R. (1987). Semantic Database Modeling: Survey, Applications, and Research Issues. *ACM Comput. Surv.*, *19*(3), 201–260. doi:10.1145/45072.45073

Humfrey, N. J. (2014). RedStore. Retrieved April 08, 2014, from http://www.aelius.com/njh/redstore/

Hunton, J. E., Libby, R., & Mazza, C. L. (2006). Financial Reporting Transparency and Earnings Management. *The Accounting Review*, *81*(1), 135–157. doi:10.2308/accr.2006.81.1.135

Hutt, K. (2005). A comparison of RDF query languages. In *Proc. of 21th Computer Science Seminar, Hartfort, Connecticut* (pp. 1–7).

IASB-IFRS. (2014). *IFRS-Who we are and what we do* (pp. 1–7). Retrieved from http://www.ifrs.org/The-organisation/Documents/WhoWeAre_JAN-2014_ENG.PDF

IEEE. (1990). IEEE Standard glossary of software engineering terminology. *Office*. New York, NY, USA: IEEE Computer Society. doi:10.1109/IEEESTD.1990.101064

Iordanov, B. (2010). HyperGraphDB: a generalized graph database. In *Web-Age Information Management* (pp. 25–36). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/978-3-642-16720-1_3

IOSCO. (2000, May 17). IASC Standards. *IOSCO Announces Completion of Its Assessment of the Accounting Standards Issued by the International Accounting Standards Committee (IASC)*, pp. 1–6. Sydney, Australia. Retrieved from http://www.iosco.org/news/pdf/IOSCONEWS26.pdf

ISO. (1997). Country Codes - ISO 3166. *What is ISO 3166?*. Retrieved June 12, 2014, from http://www.iso.org/iso/country_codes.htm

ISO. (2008a). Currency codes - ISO 4217. *What is ISO 4217?*. Retrieved June 06, 2014, from http://www.iso.org/iso/home/standards/currency_codes.htm

ISO. (2008b). News. *Publication of ISO/IEC 29500:2008, Information technology - Document description and processing languages - Office Open XML file formats*. Retrieved May 25, 2014, from http://www.iso.org/iso/news.htm?refid=Ref1181

Jacobs, I., & Walsh, N. (2004). Architecture of the world wide web. *W3C Recommendation*. Retrieved May 01, 2013, from http://www.w3.org/TR/webarch/

Janev, V., & Vraneš, S. (2011). Applicability assessment of Semantic Web technologies. *Information Processing & Management*, *47*(4), 507–517. doi:http://dx.doi.org/10.1016/j.ipm.2010.11.002

Jiménez-Domingo, E. (2013). *Modelo de Interoperabilidad para Plataformas de Cloud Computing basado en Tecnologías del Conocimiento*. Universidad Carlos III de Madrid. Retrieved from

http://e-archivo.uc3m.es/bitstream/handle/10016/18172/tesis_enrique_jimenez_domingo_2013.pdf?sequence=1

Kalyanpur, A., Parsia, B., & Hendler, J. (2005). A Tool for Working with Web Ontologies. *International Journal on Semantic Web and Information Systems*, *1*(1), 36–49. doi:10.4018/jswis.2005010103

Kambil, A., & Ginsburg, M. (1998). Public Access Web Information Systems: Lessons from the Internet EDGAR Project. *Commun. ACM*, *41*(7), 91–97. doi:10.1145/278476.278493

Karp, P. D., Chaudhri, V. K., & Thomere, J. (1999). XOL: An XML-based ontology exchange language. July. Retrieved from http://www.ai.sri.com/~pkarp/xol/

Karvounarakis, G., Alexaki, S., Christophides, V., Plexousakis, D., & Scholl, M. (2002). RQL: A Declarative Query Language for RDF. In *Proceedings of the 11th International Conference on World Wide Web* (pp. 592–603). New York, NY, USA: ACM. doi:10.1145/511446.511524

Katz, H. (2004). *XsRQL: an XQuery-style Query Language for RDF. Online only*. Retrieved from http://lists.w3.org/Archives/Public/public-rdf-dawg/2004AprJun/0740.html

Kaufmann, E., & Bernstein, A. (2010). Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. *Web Semantics: Science, Services and Agents on the World Wide Web*, *8*(4), 377–393. doi:http://dx.doi.org/10.1016/j.websem.2010.06.001

Khondoker, M. R., & Mueller, P. (2010). Comparing ontology development tools based on an online survey.

Kifer, M., & Boley, H. (2013). RIF Overview. *W3C Working Group Note*. Retrieved June 10, 2014, from http://www.w3.org/TR/rif-overview/

Kifer, M., Lausen, G., & Wu, J. (1995). Logical Foundations of Object-oriented and Frame-based Languages. *J. ACM*, *42*(4), 741–843. doi:10.1145/210332.210335

Kimmel, P. D., Weygandt, J. J., & Kieso, D. E. (2010). *Financial accounting: tools for business decision making* (6th ed., pp. 416, 527, 700). John Wiley & Sons.

Kiryakov, A., Ognyanov, D., & Manov, D. (2005). OWLIM--a pragmatic semantic repository for OWL. In *Web Information Systems Engineering--WISE 2005 Workshops* (pp. 182–192). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/11581116_19

Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, *2*(1), 49–79.

Klas, W., & Schrefl, M. (1995). Semantic data modelling. *Metaclasses and Their Application: Data Model Tailoring and Database Integration*, 71–81. doi:10.1007/BFb0027189

Kobayashi, N., & Toyoda, T. (2008). Statistical search on the Semantic Web. *Bioinformatics (Oxford, England)*, *24*(7), 1002–10. doi:10.1093/bioinformatics/btn054

Krajewski, S. (1981). Decidability. In *Dictionary of Logic as Applied in the Study of Language* (pp. 74–78). Springer. doi:10.1007/978-94-017-1253-8_17

Laguna, J. P., & Romero, E. C. (2009). Las sociedades que utilizan las NIC/NIIF ya no tienen que reconciliar las cifras contables a us gaap. ¿está justificada esta decisión de la sec en el ámbito del sector de las telecomunicaciones? *Revista de Contabilidad*, *12*(1), 45–93. doi:10.1016/S1138-4891(09)70002-0

Landin, P. J. (1964). The Mechanical Evaluation of Expressions. *The Computer Journal*, *6*(4), 308–320. doi:10.1093/comjnl/6.4.308

Lara, R., Cantador, I., & Castells, P. (2006). XBRL taxonomies and OWL ontologies for investment funds. In *Advances in Conceptual Modeling-Theory and Practice* (pp. 271–280). Springer Berlin Heidelberg. doi:10.1007/11908883_33

Larrán-Jorge, M., & Giner, B. (2002). The use of the Internet for corporate reporting by Spanish companies. *The International Journal of Digital Accounting Research*, *2*(03), 4.

Lassila, O., & Swick, R. R. (1999). Resource Description Framework (RDF) model and syntax specification. Citeseer. Retrieved March 05, 2014, from http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

Lawson, T. (2002). *Economics and reality*. Routledge.

Lenat, D. B., & Guha, R. V. (1989). *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project* (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Leuz, C. (2003). IAS Versus U.S. GAAP: Information Asymmetry–Based Evidence from Germany's New Market. *Journal of Accounting Research*, *41*(3), 445–472. doi:10.1111/1475-679X.00112

Lev, B., & Thiagarajan, S. R. (1993). Fundamental information analysis. *Journal of Accounting Research*, *31*(2), 190–215.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *10*, 707.

Liebig, T., & Noppens, O. (2004). OntoTrack: Combining browsing and editing with reasoning and explaining for OWL Lite ontologies. In *The Semantic Web--ISWC 2004* (pp. 244–258). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/978-3-540-30475-3_18

Lindörfer, F. (2010). *Semantic Web Frameworks-Jena, Joseki, Fuseki & Pellet*. Retrieved from http://cs-wwwarchiv.cs.unibas.ch/lehre/hs10/cs341/_Downloads/Workshop/Reports/2010-HS-DIS-F_Lindoerfer-Semantic_Web_Frameworks-Report.pdf

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. T., & Camón-Herrero, J. B. (2011). Financial news semantic search engine. *Expert Systems with Applications*, *38*(12), 15565–15572. doi:10.1016/j.eswa.2011.06.003

MacGregor, R., & Bates, R. (1987). *The Loom Knowledge Representation Language.*

MacGregor, R. M. (1991). Inside the LOOM Description Classifier. *SIGART Bull.*, *2*(3), 88–92. doi:10.1145/122296.122309

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, *16*(2), 72–79.

Mann, P. S. (1995). *Introductory Statistics* (2nd ed.). John Wiley & Sons Inc.

Manola, F., Miller, E., Beckett, D., & Herman, I. (2007). RDF primer a Turtle Cersion. Retrieved from http://www.w3.org/2007/02/turtle/primer/#L1323

Manola, F., Miller, E., & McBride, B. (2014). RDF 1.1 Primer. *W3C recommendation*. Retrieved June 09, 2014, from http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

Masinter, L., Berners-Lee, T., & Fielding, R. T. (2005). Uniform Resource Identifier (URI): Generic Syntax. Retrieved from http://tools.ietf.org/html/rfc3986

Matsatsinis, N. F., Doumpos, M., & Zopounidis, C. (1997). Knowledge acquisition and representation for expert systems in the field of financial analysis. *Expert Systems with Applications*, *12*(2), 247–262. doi:10.1016/S0957-4174(96)00098-X

Mcguinness, D. L., Fikes, R., Hendler, J., & Stein, L. A. (2002). DAML+OIL: an ontology language for the Semantic Web. *Intelligent Systems, IEEE*, *17*(5), 72–80. doi:10.1109/MIS.2002.1039835

McGuinness, D. L., Fikes, R., Rice, J., & Wilder, S. (2000). The chimaera ontology environment. *AAAI/IAAI*, *2000*, 1123–1124.

McGuinness, D. L., & Van-Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, *10*(2004-03), 10. Retrieved from http://www.w3.org/TR/owl-features/

Miller, L., Seaborne, A., & Reggiori, A. (2002). Three implementations of SquishQL, a simple RDF query language. In *The Semantic Web—ISWC 2002* (pp. 423–435). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/3-540-48005-6_36

Miller, P., Styles, R., & Heath, T. (2008). Open Data Commons, a License for Open Data. *LDOW*, *369*.

Mizoguchi, R., & Ikeda, M. (1998). Towards ontology engineering. *Journal-Japanese Society for Artificial Intelligence*, *13*, 9–10.

Montero, J. M., & Fernández-Aviles, G. (2010). *Enciclopedia de economía, finanzas y negocios*. Madrid: Editorial CISS (Grupo Wolters Kluwer).

Morsey, M., Lehmann, J., Auer, S., & Ngomo, A.-C. N. (2011). DBpedia SPARQL benchmark--performance assessment with real queries on real data. In *The Semantic Web--ISWC 2011* (pp. 454–469). Springer.

Motta, E. (1998). An overview of the OCML modelling language. In *the 8th Workshop on Methods and Languages*.

Motta, E. (1999). *Reusable components for knowledge modelling: Case studies in parametric design problem solving* (Vol. 53). IOS press.

Nadkarni, P. M. (1999). The EAV/CR Model of Data Representation. *The EAV/CR Physical Data Model for Heterogeneous Scientific Databases*. Retrieved May 14, 2014, from http://ycmi.med.yale.edu/nadkarni/eav_cr_frame.htm

Nadkarni, P. M., Marenco, L., Chen, R., Skoufos, E., Shepherd, G., & Miller, P. (1999). Organization of Heterogeneous Scientific Data Using the EAV/CR Representation. *Journal of the American Medical Informatics Association*, *6*(6), 478–493. doi:10.1136/jamia.1999.0060478

Ngomo, A.-C. N., & Auer, S. (2011). LIMES: A Time-efficient Approach for Large-scale Link Discovery on the Web of Data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three* (pp. 2312–2317). AAAI Press. doi:10.5591/978-1-57735-516-8/IJCAI11-385

Nobes, C., & Parker, R. H. (2008). International harmonization of accounting. In C. Nobes & R. Parker (Eds.), *Comparative international accounting* (7th ed., pp. 72–102). Upper Saddle River, NJ: Prentice-Hall.

Norrie, M. (1994). An extended entity-relationship approach to data management in object-oriented systems. In R. Elmasri, V. Kouramajian, & B. Thalheim (Eds.), *Entity-Relationship Approach — ER '93 SE - 31* (Vol. 823, pp. 390–401). Springer Berlin Heidelberg. doi:10.1007/BFb0024382

Noy, N. F., Fergerson, R. W., & Musen, M. A. (2000). The knowledge model of Protege-2000: Combining interoperability and flexibility. In *Knowledge Engineering and Knowledge Management Methods, Models, and Tools* (pp. 17–32). Springer. doi:http://dx.doi.org/10.1007/3-540-39967-4_2

Núñez, S. M., de Andrés Suárez, J., Gayo, J. E. L., & de Pablos, P. O. (2008). A semantic based collaborative system for the interoperability of XBRL accounting information. In *Emerging technologies and information systems for the knowledge society* (pp. 593–599). Springer.

O'Riain, S., Curry, E., & Harth, A. (2012). XBRL and open data for global financial ecosystems: A linked data approach. *International Journal of Accounting Information Systems*, *13*(2), 141–162. doi:10.1016/j.accinf.2012.02.002

O'Riain, S., Harth, A., & Curry, E. (2012). Linked Data Driven Information Systems as an Enabler for Integrating Financial Data. In A. Y. Yap (Ed.), *Information Systems for Global Financial Markets: Emerging Developments and Effects* (pp. 239–270). IGI Global. doi:10.4018/978-1-61350-162-7.ch010

Ogbuji, C. (2005). Versa: Path-Based RDF Query Language. O'Reilly media - xml.com. Retrieved from http://www.xml.com/pub/a/2005/07/20/versa.html

Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*, *11*(4), 295–329. doi:http://dx.doi.org/10.1016/0165-4101(89)90017-7

Ou, J., & Penman, S. (1989). Accounting measurement, price-earnings ratio, and the information content of security prices. *Journal of Accounting Research*. Retrieved from http://www.jstor.org/stable/10.2307/2491068

Owens, A., Seaborne, A., Gibbins, N., & mc schraefel. (2008). *Clustered TDB: A Clustered Triple Store for Jena. WWW2009*. Retrieved from http://eprints.soton.ac.uk/266974/

Pace, D. K. (2000). Conceptual model development for C4ISR simulations. In *Proceedings of the 5th international command and control research and technology symposium* (pp. 24–26). Laurel, Maryland. Washington, D.C. USA.

Papakonstantinou, Y., & Vianu, V. (2000). DTD Inference for Views of XML Data. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (pp. 35–46). New York, NY, USA: ACM. doi:10.1145/335168.335173

Peckham, J., & Maryanski, F. (1988). Semantic data models. *ACM Computing Surveys (CSUR)*, *20*(3), 153–189.

Penman, S. H. S. (2009). *Financial statement analysis and security valuation* (5th ed.). New York, NY, USA: McGraw-Hill Education.

Pérez, J., Arenas, M., & Gutierrez, C. (2006). Semantics and Complexity of SPARQL. In *The Semantic Web-ISWC 2006* (pp. 30–43). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/11926078_3

Pinsker, R., & Li, S. (2008). Costs and Benefits of XBRL Adoption: Early Evidence. *Commun. ACM*, *51*(3), 47–50. doi:10.1145/1325555.1325565

Plerou, V., Gopikrishnan, P., Nunes Amaral, L. A., Meyer, M., & Stanley, H. E. (1999). Scaling of the distribution of price fluctuations of individual companies. *Phys. Rev. E*, *60*(6), 6519–6529. doi:10.1103/PhysRevE.60.6519

Plumlee, R., & Plumlee, M. (2008). Assurance on XBRL for financial reporting. *Accounting Horizons*, *22*(3), 353–368. doi:http://dx.doi.org/10.2308/acch.2008.22.3.353

Powers, S. (2003). Specialized RDF Relationships: Reification, Containers, and Collections. In *Practical rdf* (1st ed., pp. 57–82). O'Reilly Media.

Prud'hommeaux, E., Carothers, G., & Machina, L. (2014). *Rdf 1.1 turtle*. Retrieved from http://www.w3.org/TR/2014/REC-turtle-20140225/

Prud'Hommeaux, E., & Seaborne, A. (2008). SPARQL query language for RDF. *W3C recommendation*. Retrieved March 15, 2014, from http://www.w3.org/TR/rdf-sparql-query/

Pryde, C., Piechocki, M., John, C. St., Warren, P., & North, D. (2013). Units Registry - Structure 1.0. *XBRL International Inc. Recommendation*. Retrieved June 11, 2014, from http://www.xbrl.org/Specification/utr/REC-2013-11-18/utr-REC-2013-11-18-clean.html

Radzimski, M., Sánchez-Cervantes, J. L., Garcia-Crespo, A., & Temiño-Aguirre, I. (2014). Intelligent Architecture for Comparative Analysis of Public Companies Using Semantics and XBRL Data. *International Journal of Software Engineering and Knowledge Engineering*, *24*(05), 801–823. doi:10.1142/S0218194014500314

Radzimski, M., Sánchez-Cervantes, J. L., Rodríguez-González, A., Gómez-Berbís, J. M., & García-Crespo, Á. (2012). FLORA--Publishing Unstructured Financial Information in the Linked Open Data Cloud. In *International Workshop on Finance and Economics on the Semantic Web (FEOSW 2012)* (pp. 27–28).

Raymond, E. S. (1996). *The new hacker's dictionary*. Mit Press.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross Validation. In M. T. Özsu & L. Liu (Eds.), *Encyclopedia of Database Systems*. Springer. Retrieved from http://www.public.asu.edu/~ltang9/papers/ency-cross-validation.pdf

Reynolds, D. (2011). Payments Ontology: Reference. *Guide to the Payments Ontology*. Retrieved August 25, 2014, from http://data.gov.uk/resources/payments

Richards, J., Smith, B., & Saeedi, A. (2006). An introduction to XBRL. *Available at SSRN 1007570*. doi:http://dx.doi.org/10.2139/ssrn.1007570

Robinson, S. (2006). Conceptual Modeling for Simulation: Issues and Research Requirements. In *Proceedings of the 38th Conference on Winter Simulation* (pp. 792–800). Winter Simulation Conference. Retrieved from http://dl.acm.org/citation.cfm?id=1218112.1218259

Rodríguez-González, A., Colomo-Palacios, R., Guldris-Iglesias, F., Gómez-Berbís, J., & García-Crespo, A. (2012). FAST: Fundamental Analysis Support for Financial Statements. Using semantics for trading recommendations. *Information Systems Frontiers*, *14*(5), 999–1017. doi:10.1007/s10796-011-9321-1

Rodríguez-González, A., García-Crespo, Á., Colomo-Palacios, R., Guldrís Iglesias, F., & Gómez-Berbís, J. M. (2011). CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator. *Expert Systems with Applications*, *38*(9), 11489–11500. doi:10.1016/j.eswa.2011.03.023

Rohloff, K., Dean, M., Emmons, I., Ryder, D., & Sumner, J. (2007). An evaluation of triple-store technologies for large data stores. In *On the Move to Meaningful Internet*

Systems 2007: OTM 2007 Workshops (pp. 1105–1114). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/978-3-540-76890-6_38

Roohani, S., Furusho, Y., & Koizumi, M. (2009). XBRL: Improving transparency and monitoring functions of corporate governance. *International Journal of Disclosure and Governance*, *6*(4), 355–369. Retrieved from http://dx.doi.org/10.1057/jdg.2009.17

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., Lorensen, W. E., & others. (1991). *Object-oriented modeling and design* (Vol. 199). Prentice hall Englewood Cliffs (NJ).

Sánchez-Cervantes, J. L., Hernández-Chan, G. S., Radzimski, M., Gómez-Berbís, J. M., & García-Crespo, Á. (2013). Discovering and Linking Financial Data on the Web. In *DATA ANALYTICS 2013, The Second International Conference on Data Analytics* (pp. 36–40).

Sauermann, L., Cyganiak, R., & Völkel, M. (2008). Cool URIs for the semantic web. *W3C Working Group Note*. Retrieved April 12, 2014, from http://www.w3.org/TR/cooluris/

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Lecture Notes in Statistics-New York-Springer Verlag*, 149–172.

Schmachtenberg, M., Paulheim, H., & Bizer, C. (2014). Adoption of Linked Data Best Practices in Different Topical Domains. In *The 13th International Semantic Web Conference (ISWC2014)* (pp. 1–16). Riva del Garda-Trentino, Italy. Retrieved from http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/

Schmid, H. A., & Swenson, J. R. (1975). On the Semantics of the Relational Data Model. In *Proceedings of the 1975 ACM SIGMOD International Conference on Management of Data* (pp. 211–223). New York, NY, USA: ACM. doi:10.1145/500080.500110

Schnase, J. L., Leggett, J. J., Hicks, D. L., & Szabo, R. L. (1993). Semantic Data Modeling of Hypermedia Associations. *ACM Trans. Inf. Syst.*, *11*(1), 27–50. doi:10.1145/151480.151521

Schumacher, M. (2003). *Security engineering with patterns: origins, theoretical models, and new applications* (Vol. 2754, pp. 29–44). Springer. doi:http://dx.doi.org/10.1007/978-3-540-45180-8_3

Schwarz, G., & others. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Seaborne, A. (2004). RDQL-A query language for rdf. *W3C Member submission*. Retrieved March 15, 2014, from http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/

Segaran, T., Evans, C., & Taylor, J. (2009). *Programming the semantic web*. O'Reilly Media.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web Revisited. *Intelligent Systems, IEEE*, *21*(3), 96–101. doi:10.1109/MIS.2006.62

Shamsfard, M., & Barforoush, A. A. (2004). Learning ontologies from natural language texts. *International Journal of Human-Computer Studies*, *60*(1), 17–63. doi:10.1016/j.ijhcs.2003.08.001

Sheth, A., Bertram, C., Avant, D., Hammond, B., Kochut, K., & Warke, Y. (2002). Managing semantic content for the Web. *Internet Computing, IEEE*, *6*(4), 80–87.

Sheth, A. P., & Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR)*, *22*(3), 183–236.

Shin, R. Y. (2003). XBRL, Financial Reporting, and Auditing. *The CPA Journal*, *73*(12), 61.

Silberschatz, A., Sudarshan, S., & Korth, H. F. (2002). *Fundamentos de bases de datos* (4th ed., pp. 5–7). Madrid Esp.: McGraw-Hill Inc.

Sinnett, W. M. (2011). SEC Reporting and the Impact of XBRL: 2011 Survey. *XBRL. Financial Executives Research Foundation (FERF)*. Retrieved October 10, 2013, from http://www.financialexecutives.org/ferf/download/2011 Final/2011-031.pdf

Sintek, M., & Decker, S. (2002). TRIPLE—A query, inference, and transformation language for the semantic web. In *The Semantic Web—ISWC 2002* (pp. 364–378). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/3-540-48005-6_28

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, *5*(2), 51–53. doi:10.1016/j.websem.2007.03.004

Smith, J. M., & Smith, D. C. P. (1977). Database Abstractions: Aggregation and Generalization. *ACM Trans. Database Syst.*, *2*(2), 105–133. doi:10.1145/320544.320546

Steedman, D. (1993). *Abstract syntax notation one (ASN. 1): the tutorial and reference*. Technology appraisals. Retrieved from http://www.bgbm.fu-berlin.de/tdwg/acc/Documents/asn1gloss.htm

Street, L. D., Gray, S. J., & Bryant, S. M. (1999). Acceptance and Observance of International Accounting Standards: An Empirical Study of Companies Claiming to Comply with IASs. *The International Journal of Accounting*, *34*(1), 11–48. doi:10.1016/S0020-7063(99)80002-8

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1–2), 161–197. doi:http://dx.doi.org/10.1016/S0169-023X(97)00056-6

Stümpert, T. (2008). Extracting Financial Data from SEC Filings for US GAAP Accountants. In D. Seese, C. Weinhardt, & F. Schlottmann (Eds.), *Handbook on Information Technology in Finance SE - 16* (pp. 357–375). Springer Berlin Heidelberg. doi:10.1007/978-3-540-49487-4_16

Sure, Y., Angele, J., & Staab, S. (2002). OntoEdit: Guiding ontology development by methodology and inferencing. In *On the Move to Meaningful Internet Systems 2002: CoopIS,*

DOA, and ODBASE (pp. 1205–1222). Springer Berlin Heidelberg. doi:http://dx.doi.org/10.1007/3-540-36124-3_76

Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). *OntoEdit: Collaborative ontology development for the semantic web* (pp. 221–235). Springer. doi:http://dx.doi.org/10.1007/3-540-48005-6_18

Swartout, B., Patil, R., Knight, K., & Russ, T. (1996a). Ontosaurus: a tool for browsing and editing ontologies. In *9th Banff Knowledge Aquisition for KNowledge-based systems Workshop*.

Swartout, B., Patil, R., Knight, K., & Russ, T. (1996b). Toward distributed use of large-scale ontologies. In *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*.

Tamayo, M. (2004). *El proceso de la investigaci{ó}n cient{í}fica* (4th ed., p. 156). Editorial Limusa.

Taylor, R. W., & Frank, R. L. (1976). CODASYL Data-Base Management Systems. *ACM Comput. Surv.*, *8*(1), 67–103. doi:10.1145/356662.356666

Teorey, T. J. (1990). *Database Modeling and Design: The Entity-relationship Approach*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ter Horst, H. J. (2005). Completeness, decidability and complexity of entailment for {RDF} Schema and a semantic extension involving the {OWL} vocabulary. *Web Semantics: Science, Services and Agents on the World Wide Web*, *3*(2–3), 79–115. doi:http://dx.doi.org/10.1016/j.websem.2005.06.001

Thompson, B., & Personick, M. (2009). Bigdata: the semantic web on an open source cloud. In *International Semantic Web Conference*.

Thompson, H. (2012). XML schema part 1: structures second edition. *2004-10. Http://www. w3. org/TR/2004/REC-Xmlschema-1-20041028*. Retrieved from http://www.w3.org/TR/xmlschema11-1/

Ting, K. (2010). Precision and Recall. In C. Sammut & G. Webb (Eds.), *Encyclopedia of Machine Learning SE - 652* (p. 781). Springer US. doi:10.1007/978-0-387-30164-8_652

Trochim, W. M. K. (2006). Descriptive statistics. *Research Methods Knowledge Base*. Research Methods of Knowledge. Retrieved September 21, 2014, from http://www.socialresearchmethods.net/kb/statdesc.php

Tsichritzis, D. C., & Lochovsky, F. H. (1976). Hierarchical Data-Base Management: A Survey. *ACM Comput. Surv.*, *8*(1), 105–123. doi:10.1145/356662.356667

U.S. SEC. (2013). *Filer Manual -Volume II EDGAR Filing* (pp. 3–18). Retrieved from http://www.sec.gov/info/edgar/edgarfm-vol1-v15.pdf

U.S. SEC. (2014a). Filing Detail. *APPLE INC-Filing Detail*.

U.S. SEC. (2014b). View Filing Data. *APPLE INC-View Filing Data*. Retrieved September 10, 2014, from http://www.sec.gov/cgi-bin/viewer?action=view&cik=320193&accession_number=0001193125-14-157311&xbrl_type=v

Van-Tendeloo, B., & Vanstraelen, A. (2005). Earnings management under German GAAP versus IFRS. *European Accounting Review*, *14*(1), 155–180. doi:10.1080/0963818042000338988

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., & Studer, R. (2006). Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 585–594). New York, NY, USA: ACM. doi:10.1145/1135777.1135863

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009a). Discovering and Maintaining Links on the Web of Data. In A. Bernstein, D. Karger, T. Heath, L. Feigenbaum, D. Maynard, E. Motta, & K. Thirunarayan (Eds.), *The Semantic Web - ISWC 2009 SE - 41* (Vol. 5823, pp. 650–665). Springer Berlin Heidelberg. doi:10.1007/978-3-642-04930-9_41

Volz, J., Bizer, C., Gaedke, M., & Kobilarov, G. (2009b). Silk-A Link Discovery Framework for the Web of Data. In *LDOW*.

W3C. (2011a). Benefits of the Linked Data Approach. *Benefits*. Retrieved July 14, 2014, from http://www.w3.org/2005/Incubator/lld/wiki/Benefits

W3C. (2011b). XML in 10 points. *XML in 10 points*. Retrieved May 25, 2014, from http://www.w3.org/XML/1999/XML-in-10-points.html.en

W3C. (2013). SPARQL 1.1 Overview. *W3C Recommendation*. World Wide Web Consortium. Retrieved March 15, 2014, from http://www.w3.org/TR/sparql11-overview/

W3C-Group. (2012). OWL 2 Web Ontology Language Document Overview, W3C Recommendation, Dec, 2012. Retrieved from http://www.w3.org/TR/owl2-overview/#ref-owl-2-profiles

Wilkinson, K., Sayers, C., Kuno, H. A., & Reynolds, D. (2003). Efficient RDF Storage and Retrieval in Jena2. In *SWDB* (Vol. 3, pp. 131–150).

Williams, J. R., Haka, S. F., Bettner, M. S., & Carcello, J. V. (2005). *Financial and managerial accounting*. China Machine Press.

Winnenburg, R., & Bodenreider, O. (2014). Desiderata for an authoritative Representation of MeSH in RDF. Retrieved from http://mor1.nlm.nih.gov/pubs/pdf/2014-amia-rw.pdf

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., … Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, *34*(Database issue), D668–72. doi:10.1093/nar/gkj067

Woelk, D., Kim, W., & Luther, W. (1990). Readings in Object-oriented Database Systems. In S. B. Zdonik & D. Maier (Eds.), (pp. 592–606). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Retrieved from http://dl.acm.org/citation.cfm?id=93490.94149

Worboys, M. F., Hearnshaw, H. M., & Maguire, D. J. (1990). Object-oriented data modelling for spatial databases. *International Journal of Geographical Information Systems*, *4*(4), 369–383. doi:10.1080/02693799008941553

Worldbank. (2012). World Bank Linked Data. Retrieved May 22, 2014, from http://worldbank.270a.info/.html

Wu, J., & Vasarhelyi, M. (2004). XBRL: A New Tool For Electronic Financial Reporting. In M. Anandarajan, A. Anandarajan, & C. Srinivasan (Eds.), *Business Intelligence Techniques SE - 5* (pp. 73–92). Springer Berlin Heidelberg. doi:10.1007/978-3-540-24700-5_5

XBRL-España. (2005). XBRL y el plan de convergencia. *3*. Retrieved from http://www.xbrl.es/boletin/03/plan.html

XBRL-España. (2006). *Libro blanco XBRL* (pp. 18–23). Retrieved from http://www.xbrl.es/downloads/libros/Libro_Blanco.pdf

XBRL-International. (2013). Units Registry. *Units Registry Specification*. Retrieved June 11, 2014, from http://xbrl.org/utr/utr.xml

Yadagiri, N., & Ramesh, P. (2013). Semantic Web and the Libraries: An Overview. *International Journal of Library Science™*, *7*(1), 80–94.

Zhu, H., & Madnick, S. E. (2007). Semantic integration approach to efficient business data supply chain: Integration approach to interoperable XBRL. *MIT Sloan School of Management Research Paper Series*. doi:http://dx.doi.org/10.2139/ssrn.1075711

Zhu, J., & Wang, Y. (2011). Research on typical parsing models for XBRL taxonomy. In *System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2011 International Conference on* (Vol. 2, pp. 13–16). doi:10.1109/ICSSEM.2011.6081257

Zisman, A. (2000). An overview of XML. *Computing Control Engineering Journal*, *11*(4), 165–167. doi:10.1049/cce:20000405