



**UNIVERSIDAD DE CASTILLA-LA MANCHA
ESCUELA SUPERIOR DE INFORMÁTICA**

**MÁSTER UNIVERSITARIO
EN INGENIERÍA INFORMÁTICA**

TRABAJO FIN DE MÁSTER

**SparkDQ: A Spark Stack to Support Data
Quality Context-Aware Assessment of Semantic
Triples**

Raúl Reguillo Carmona

Octubre, 2017

**SPARKDQ: A SPARK STACK TO SUPPORT DATA QUALITY
CONTEXT-AWARE ASSESSMENT OF SEMANTIC TRIPLES**



UNIVERSIDAD DE CASTILLA-LA MANCHA
ESCUELA SUPERIOR DE INFORMÁTICA

TRABAJO FIN DE MÁSTER

**SparkDQ: A Spark Stack to Support Data
Quality Context-Aware Assessment of Semantic
Triples**

Autor: Raúl Reguillo Carmona

Tutor: Ismael Caballero Muñoz-Reja

Cotutor: Bibiano Rivas García

Octubre, 2017



UNIVERSIDAD DE CASTILLA-LA MANCHA

ESCUELA SUPERIOR DE INFORMÁTICA

TRABAJO FIN DE MÁSTER

**SparkDQ: A Spark Stack to Support Data
Quality Context-Aware Assessment of Semantic
Triples**

Fdo.: Raúl Reguillo Carmona Fdo.: Ismael Caballero Muñoz-Reja

Octubre, 2017

Raúl Reguillo Carmona

Ciudad Real – Spain

E-mail: raul.reguillo@alu.uclm.es

Teléfono: 638 175 608

© 2017 Raúl Reguillo Carmona

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

Se permite la copia, distribución y/o modificación de este documento bajo los términos de la Licencia de Documentación Libre GNU, versión 1.3 o cualquier versión posterior publicada por la *Free Software Foundation*; sin secciones invariantes. Una copia de esta licencia esta incluida en el apéndice titulado «GNU Free Documentation License».

Muchos de los nombres usados por las compañías para diferenciar sus productos y servicios son reclamados como marcas registradas. Allí donde estos nombres aparezcan en este documento, y cuando el autor haya sido informado de esas marcas registradas, los nombres estarán escritos en mayúsculas o como nombres propios.

TRIBUNAL:

Presidente:

Vocal:

Secretario:

FECHA DE DEFENSA:

CALIFICACIÓN:

PRESIDENTE

VOCAL

SECRETARIO

Fdo.:

Fdo.:

Fdo.:

Resumen

La Web Semántica es un tipo de web cuyo contenido puede ser procesado tanto por máquinas de forma automática como por personas. Esto es posible debido a una serie de avances significativos en la representación del conocimiento durante los últimos años. Esta representación del conocimiento se basa en el concepto de *triplas* (sujeto, predicado, objeto) que establecen relaciones entre los distintos datos. Estas relaciones a su vez se representan utilizando lenguajes y especificaciones desarrollados para tal fin, como son Framework de Descripción de Recursos (RDF) y Lenguaje de Ontología Web (OWL).

Paralelamente surge el paradigma de Linked Data, que se puede definir como el uso del modelo web para publicar datos estructurados de manera que puedan ser fácilmente consumidos y enlazados con otros. Aprovechando la capacidad semántica obtenida mediante la representación en triplas y la capacidad de poder interconectar los datos de esta manera, se pueden tener volúmenes significativos de datos interrelacionados sobre los cuales aplicar razonamiento automático, consultas o cualquier otra operación como parte del procesamiento automático anteriormente citado.

Por otra parte, tener en cuenta la calidad de los datos resulta algo imprescindible en el desarrollo con éxito de cualquier tarea. Se puede encontrar una gran cantidad de *frameworks* (Jena, Sesame, ...) con los que elaborar aplicaciones semánticas, pero actualmente ninguno de ellos incorpora primitivas para dar soporte a las operaciones de gestión de calidad de datos que están siendo utilizados. Dada la importancia de la calidad de datos y la carencia de dichas utilidades en los frameworks anteriormente citados, es patente la necesidad de elaborar mecanismos que permitan llevar a cabo evaluaciones de calidad de los datos usados en aplicaciones de tecnología semántica en general y Web Semántica en particular.

Así pues el objetivo del Proyecto Fin de Carrera consistió en la elaboración de una extensión de un framework de desarrollo de aplicaciones de Web Semántica y Linked Data: Apache Jena. La finalidad de dicha extensión fue, dado un conjunto de Linked Data, dar soporte a primitivas de medición de calidad de datos teniendo en cuenta el contexto en el que dichos datos deben ser considerados.

El objetivo principal de una posible iteración sobre este proyecto fue extender estas funcionalidades en el contexto Big Data, debido a la falta de eficiencia encontrada cuando las métricas calculadas ganaban complejidad e involucraban un número considerable de triplas.

De esta manera, el objetivo principal de este Trabajo Fin de Máster es el desarrollo de un stack para un framework de procesamiento de datos en entornos distribuidos, Apache Spark, de manera que se pueda suplir la necesidad de evaluación de calidad de datos enlazados en entornos Big Data.

Abstract

The Semantic Web is a type of Web which content can be processed automatically. This is possible due to a number of significant advances in knowledge representation in recent years. Knowledge in the Semantic Web is represented by triples (subject, predicate, object) that establish relations between different data using languages and specifications developed for this purpose, such as Resource Definition Language (RDF) and Ontology Web Language (OWL).

Meanwhile, Linked Data paradigm appears, which can be defined as the use of the Web model to publish structured data so it can be easily consumed and linked to other data. Considering the semantic capacity obtained by triples representation and the ability to interconnect data thus, users could have significant volumes of interrelated data where apply automated reasoning, inquiries or any other operation.

Moreover, it should be considered data quality as a need in the successful perform of any task. People can find a lot of frameworks (Jena, Sesame, . . .) to develop semantic application, but currently none of them incorporates basic operations for assessing quality of data in use. Many authors consider that this part has not yet been taken into account or has been received little attention. Therefore, there is a need to develop mechanisms to carry out quality assessments of the data used in semantic applications.

Thus, the objective of the original project involved the development of an extension for a Semantic Web and Linked Data framework: Apache Jena. The purpose of this extension was, having a set of Linked Data, to support a set of data quality measures taking into account the context in which these data should be considered.

The main target of a possible next iteration over the project was to extend this functionalities on a Big Data context, due to lack of efficiency found when the metrics calculation get complex and involves an increasing number of triples.

Thus, the goal of this project is the development of a stack for a distributed processing framework, Apache Spark, in order to cover the need of Data Quality Assessment of Linked Data in a Big Data environment.

Listado de acrónimos

AAA	<i>Anyone can say Anything about Any topic</i>
API	Application Programming Interface
CSV	Comma Separated Values
DQ	Calidad de Datos
DQD	Dimensión de Calidad de Datos
DOAP	Description Of A Project
ER	Entidad-Interrelación
FOAF	Friend Of A Friend
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ISO	International Organization for Standarization
LD	Datos Enlazados
LOD	Datos Enlazados Abiertos
OD	Datos Abiertos
OWL	Lenguaje de Ontología Web
PFC	Proyecto Fin de Carrera
RDF	Framework de Descripción de Recursos
RQL	Relationship Query Language
SGML	Standard Generalized Markup Language
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TDB	Triple Data Base
TFM	Trabajo Fin de Máster
URI	Identificador Uniforme de Recurso
URL	Localizador Uniforme de Recurso

0.

W3C Consorcio para la World Wide Web

XML Lenguaje de Marcas Extensible

Agradecimientos

Escribe aquí algunos chascarrillos simpáticos. Haz buen uso de todos tus recursos literarios porque probablemente será la única página que lean tus amigos y familiares. Debería caber en esta página (esta cara de la hoja).

Juan¹

¹Sí, los agradecimientos se firman

A alguien querido y/o respetado

Índice general

Resumen	V
Abstract	VII
Listado de acrónimos	IX
Agradecimientos	XI
Índice general	XV
Índice de cuadros	XVII
Índice de figuras	XIX
Índice de listados	XXI
Listado de acrónimos	XXIII
1. Introducción	1
1.1. Contexto del proyecto	1
1.2. Evolución respecto del PFC	3
1.3. Logro de competencias planteadas	4
1.4. Estructura del documento	4
2. Objetivos	7
2.1. Objetivos principales	7
2.2. Objetivos parciales	7
3. Estado del Arte	9
3.1. Web Semántica	9
3.1.1. Concepto	9
3.1.2. Terminología de Web Semántica	10

3.1.3. Estándares	11
3.1.4. Arquitectura	13
3.1.5. Ontologías	15
3.1.6. Vocabularios	17
3.1.7. Frameworks para el desarrollo de aplicaciones de Web Semántica .	19
3.2. Datos Enlazados (LD)	23
3.2.1. Definición de LD	23
3.2.2. Datos Abiertos (OD) y Datos Enlazados Abiertos (LOD)	23
3.2.3. LOD en la actualidad	25
3.3. Calidad de Datos	25
3.3.1. Dimensiones de Calidad de Datos	27
3.3.2. Completeness	27
3.3.3. Accessibility	28
3.4. Calidad de Datos en LD	28
3.5. Big Data	31
3.5.1. Concepto	31
3.5.2. Desafíos	32
3.5.3. Frameworks	33
3.5.4. Datos Semánticos en entornos Big Data	35
4. Método de Trabajo	37
5. Resultados	39
6. Conclusiones	41
6.1. Conclusiones	41
6.2. Propuestas de trabajos futuros	41
6.3. Publicaciones	41
6.4. Opinión personal	41
A. Ejemplo de anexo	45
Referencias	47

Índice de cuadros

3.1. Comparativa de frameworks de Web Semántica. Extraída de [W3Cc]	22
3.2. Categorías y dimensiones de DQ	27

Índice de figuras

3.1. Esquema de tripla	12
3.2. Grafo basado en triplas. Extraído de [JEN14]	13
3.3. Tecnologías de Web Semántica	14
3.4. Arquitectura de la Web Semántica. Extraído de [San11]	15
3.5. Clasificación de Ontologías. Extraído de [Gua98]	17
3.6. Ejemplo de FOAF. Extraído de [JEN14]	19
3.7. Diagrama de la arquitectura de Jena. Extraída de [JEN14]	21
3.8. Razonamiento en Jena. Extraída de [JEN14]	22
3.9. Esquema LD DBpedia	24
3.10. Niveles de granularidad en la Web Semántica. Extraído de [DFP ⁺ 05]	24
3.11. Bus Gijón App	26
3.12. Bus Gurú App	26
3.13. 5vs del Big Data	32
3.14. MapReduce: ejecución [DG04]	33

Índice de listados

3.1. Ejemplo de FOAF	19
3.2. Consulta SPARQL para identificación de literales perdidos (I)	29
3.3. Consulta SPARQL para identificación de literales perdidos (y II)	29

Listado de acrónimos

AAA	<i>Anyone can say Anything about Any topic</i>
API	Application Programming Interface
CSV	Comma Separated Values
DQ	Calidad de Datos
DQD	Dimensión de Calidad de Datos
DOAP	Description Of A Project
ER	Entidad-Interrelación
FOAF	Friend Of A Friend
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ISO	International Organization for Standarization
LD	Datos Enlazados
LOD	Datos Enlazados Abiertos
OD	Datos Abiertos
OWL	Lenguaje de Ontología Web
PFC	Proyecto Fin de Carrera
RDF	Framework de Descripción de Recursos
RQL	Relationship Query Language
SGML	Standard Generalized Markup Language
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TDB	Triple Data Base
TFM	Trabajo Fin de Máster
URI	Identificador Uniforme de Recurso
URL	Localizador Uniforme de Recurso

0.

W3C Consorcio para la World Wide Web

XML Lenguaje de Marcas Extensible

Capítulo 1

Introducción

Contexto del proyecto

EN [BLHL01] se define la Web Semántica como un tipo de Web cuyo contenido resulta procesable de manera automática. Para conseguir un contenido procesable se debe antes llevar a cabo una representación o formalización del conocimiento que se pretende exponer. Esto ha sido posible gracias al concepto de *tripla* que no es más que la relación, expresada en un lenguaje formal, existente entre un sujeto y un objeto a través de un predicado que los relaciona. De esta manera, si se desea expresar formalmente una sentencia respecto de cualquier concepto bastará con establecer un sujeto o recurso (el propio concepto) una relación o propiedad que se desee modelar, y un objeto que a su vez puede ser un nuevo recurso de manera que agrupando conjuntos de triplas se logra definir conceptos completos de manera formal.

Durante los últimos años se han desarrollado lenguajes como RDF y OWL, y especificaciones, con el fin de dar soporte a la interoperabilidad semántica [SBLH06]. En 1997 el Consorcio para la World Wide Web (W3C) [W3Ca] define la primera especificación de RDF lo que sentará los cimientos de la Web Semántica: el objetivo de RDF será aportar una descripción semántica del conocimiento en la Web [SBLH06] como lenguaje de definición de recursos, haciendo posible trabajar con la totalidad de la Web como un conjunto de recursos y sus interrelaciones.

Posteriormente surge el paradigma Linked Data [BL09], que se puede definir como el uso del modelo Web para publicar datos estructurados de manera que puedan ser fácilmente consumidos y combinados con otros [BHBL09]. Valiéndose del concepto de Identificador Uniforme de Recurso (URI), se consiguen identificar recursos en la red de forma unívoca y así poder enlazarlos sin ambigüedad. Linked Data aprovecha la potencia descriptiva de RDF para facilitar el procesado y razonamiento automático sobre grandes conjuntos de datos. En [BL09] se enumera una serie de principios que deben cumplir los datos para ser considerados Linked Data:

1. Usar URI como identificadores de los recursos publicados en la Web.
2. Usar las URL de estas URI para que la gente pueda localizar y consultar estos recursos.

1. INTRODUCCIÓN

3. Proporcionar información útil cuando la URI sea desreferenciada.
4. Incluir vínculos a otras URI relacionadas con los datos en el recurso.

Por otra parte, el concepto de *Calidad* puede aplicarse a cualquier tarea u objeto. Es importante matizar para el contexto del proyecto qué se entiende por calidad y más concretamente por calidad de los datos (DQ).

Existen varias definiciones de calidad que conviene considerar a la hora de abordar el proyecto. En primer lugar, se va a exponer la definición que la Real Academia Española de la lengua otorga a este término:

Propiedad o conjunto de propiedades inherentes a algo que permiten juzgar su valor.

Sin embargo este concepto ha ido variando con el tiempo, adaptándose a las necesidades y a los procesos. Relacionando este concepto con la *fabricación* de un determinado bien, en [ISH85] se define la calidad como:

Un producto tiene calidad cuando es desarrollado, diseñado y mantenido de la forma más económica, útil y satisfactoria para el consumidor.

Autores como [CMC08] o [FH10] dejan constancia de cómo la calidad de los datos resulta imprescindible en el desarrollo con éxito de cualquier tarea o toma de decisión. Un nivel inadecuado de calidad en los datos puede tener impactos sustanciales en ámbitos sociales y económicos. Las empresas están mejorando la calidad de los datos con enfoques y herramientas prácticas [WS96].

En el ámbito de las tecnologías semánticas, la comunidad ha prestado poca atención a la calidad de los datos que han sido representados mediante la utilización de tecnologías semánticas [FH10]. Existen en la actualidad un número considerable de frameworks disponibles para el desarrollo de aplicaciones basadas en el uso de tecnologías semánticas, no obstante, ninguno de ellos incluye funcionalidades específicas que permitan procesar una evaluación del nivel de calidad de los datos que se están usando.

Puesto que ningún framework de desarrollo de Web Semántica facilita este tipo de operaciones de medición, existe una necesidad de elaborar mecanismos que permitan llevar a cabo evaluaciones de calidad de los datos usados en aplicaciones de tecnologías semánticas.

Con la irrupción de las tecnologías Big Data y dadas las necesidades que existen de procesamiento distribuido sobre grandes volúmenes de información, encontramos que el problema de la evaluación de la calidad de los datos semánticos se extiende a este ámbito, no pudiendo encontrar ningún framework que permita trabajar la calidad de los datos semánticos de manera distribuida.

Considerando la no disponibilidad de operaciones de evaluación de calidad de datos en los frameworks de desarrollo de tecnologías semánticas, el objetivo del Trabajo Fin de Máster (TFM) consiste en el desarrollo de una capa tecnológica (stack) para uno de estos frame-

works. La finalidad de dicho stack será la de diseñar, implementar y poner a disposición de la comunidad un conjunto de primitivas de evaluación de calidad de datos para uno de los frameworks de desarrollo de aplicaciones de tecnología semántica más importantes, Apache Jena, considerando el contexto en el que los datos deben ser usados mediante el uso de reglas que lo modelen y sean procesables.

Evolución respecto del PFC

Tras llevar a cabo un estudio acerca del estado actual de los frameworks de desarrollo de Web Semántica y Linked Data, durante el desarrollo del Proyecto Fin de Carrera (PFC) se escogió uno de ellos basado en Java, **Apache Jena**, y se extendió con la finalidad de ofrecer estos mecanismos para la evaluación de la calidad de datos semánticos.

Paralelamente, durante los últimos años han tenido lugar una serie de cambios muy significativos en la industria de las tecnologías de la información, tales como el nacimiento del Internet de las Cosas, Computación en la Nube y la aparición de las Redes Sociales.

El desarrollo y explotación de estas tecnologías ha propiciado que la cantidad de los datos que se generan por unidad de tiempo se haya incrementado a una velocidad sin precedentes, trayendo consigo la necesidad de cómputo de tales volúmenes de información heterogénea, en tiempos razonables y de manera escalable [DG04].

Esta necesidad ha propiciado la aparición de paradigmas, tecnologías y numerosos frameworks con los que abordar esta problemática desde diferentes puntos de vista (procesamiento por lotes, en tiempo real, extracción, transformación y carga de datos, seguridad, . . .), agrupándose en ecosistemas tal y como pueda ser el caso del **ecosistema Hadoop** [HAD].

Apache Spark [SPAa] es un framework de procesamiento distribuido, perteneciente al ecosistema Hadoop, que en los últimos años se ha convertido en el estándar *de facto* a la hora de abordar problemas de procesamiento de grandes volúmenes de información en tiempos razonables, obteniendo unos resultados de rendimiento varios órdenes de magnitud superior frente a modelos y tecnologías anteriores, debido al procesamiento en memoria.

Mientras que el proyecto de Jena ha ido evolucionando para permitir el procesamiento de datos semánticos de manera centralizada y habiendo estudiado el estado actual de los frameworks de procesamiento distribuido, se concluye que ninguno de ellos ofrece primitivas para el procesamiento de datos semánticos en distribuido y por lo tanto, tampoco mecanismos para llevar a cabo cálculos de métricas de calidad sobre datos semánticos en entornos distribuidos.

Debido a la necesidad de incluir primitivas de evaluación de calidad de los datos a la hora de desarrollar cualquier tarea o toma de decisión y puesto que el ritmo de crecimiento de los datos dificulta su análisis en tiempos razonables, se establece como objetivo principal de este Trabajo Fin de Máster ofrecer un stack tecnológico basado en un framework de proce-

1. INTRODUCCIÓN

samiento distribuido para que permita llevar a cabo análisis de calidad de datos semánticos en entornos Big Data.

Para el presente Trabajo fin de Máster, será **Apache Spark** el framework sobre el cual se construirá dicho stack para dar soporte a primitivas de evaluación de calidad de los datos de una manera distribuida, escalable y tolerante a fallos.

Como incremento respecto al PFC, cabe destacar los siguientes aspectos:

1. Elaboración de un stack tecnológico basado en un framework de desarrollo distribuido, **Apache Spark**, para permitir el procesado de datos RDF: **SparkRDF**. Este stack estará basado en el framework **Apache Jena** para el procesamiento de datos en formatos semánticos y en **Apache Spark** para ofrecer primitivas de procesamiento en distribuido.
2. Extensión de dicho stack para permitir evaluación de calidad de datos semánticos de manera distribuida, **SparkDQ**.
3. Elección o desarrollo de una ontología que describa evaluaciones de calidad de los datos.
4. Desarrollo de primitivas de razonamiento sobre evaluaciones de calidad.
5. Incremento del rendimiento a la hora de calcular métricas sobre grafos de datos enlazados.
6. Estudio y comparativa entre resultados de **JenaDQ** y **SparkDQ** con el fin de establecer unos criterios de mejora en términos de tiempo y eficiencia a la hora de llevar a cabo estos cálculos de métricas de calidad.
7. Elaboración de una prueba de concepto de extracción de grandes volúmenes de datos, evaluación y presentación de las métricas.

Logro de competencias planteadas

TODO

Pues ha ido flaman.

Estructura del documento

La forma en la que se ha organizado el presente documento se expone a continuación:

Capítulo 1: Introducción

En este apartado se introduce el problema y se plantea el trabajo en base a los objetivos fijados.

Capítulo 2: Objetivos

En este capítulo se describe el principal objetivo del TFM así como un conjunto de

objetivos parciales que es necesario alcanzar para cubrir el objetivo final.

Capítulo 3: Estado del Arte

Esta sección introducirá los conceptos más importantes tratados en el documento y el proyecto, acompañado de conceptos y definiciones. Incluirá referencias al ámbito de la calidad de los datos, la Web Semántica y las tecnologías asociadas a ésta.

Capítulo ??: ??

Se expondrá detalladamente cómo se ha utilizado la metodología de desarrollo elegida para la realización del TFM.

Capítulo 5: Resultados

En este capítulo se expondrán los resultados obtenidos durante las diferentes iteraciones que se han llevado a cabo durante el desarrollo del TFM.

Capítulo 6: Conclusiones

En este capítulo se expondrán las conclusiones que se han obtenido tras la realización de este TFM además de una serie de propuestas de trabajo futuro.

Anexo ??: ??

Se corresponde con el listado de código de las ontologías desarrolladas para formalizar los resultados de evaluaciones de calidad de datos enlazados.

Capítulo 2

Objetivos

Objetivos principales

Los objetivos de este trabajo fin de máster (Trabajo Fin de Máster (TFM)) son:

- Desarrollar una extensión para el framework XX con el fin de dar soporte a primitivas de evaluación de calidad de datos en un entorno Big Data. Dicha extensión se generará como una API cuya finalidad será encapsular todas las primitivas desarrolladas y facilitar al usuario la utilización de dicha extensión.
 - Establecer un modelo de desarrollo distribuido para uno de estos frameworks con el fin de dar soporte a primitivas de evaluación de calidad de datos en entorno Big Data.
- Propuesta de arquitectura Big Data orientada al manejo y evaluación de datos semánticos.

Objetivos parciales

El objetivo principal implica el abordar una serie de objetivos parciales relacionados con la arquitectura y funcionalidades del sistema que se pretenden desarrollar. Dichos objetivos parciales se detallan a continuación.

O1. Representación del contexto de evaluación de calidad de datos

Implementar las reglas de negocio que representen el contexto necesario para la evaluación de calidad de datos usando la tecnología más adecuadas.

O2. Desarrollo de un vocabulario para los resultados de evaluaciones de calidad de datos

El fin de dicho vocabulario será el de permitir publicar los resultados de las evaluaciones que se lleven a cabo de manera formal y basándose en los paradigmas de la Web Semántica. Los conceptos de ontología y vocabulario serán detallados en la Sección 3.1.5.

O3. Diseño e implementación de las primitivas de evaluación de calidad de datos

Diseñar e implementar primitivas para la evaluación de calidad de los datos como una

2. OBJETIVOS

extensión del framework elegido, considerando el contexto de uso mediante el uso de reglas semánticas.

Estas primitivas se desarrollarán tomando diferentes perspectivas de la calidad de los datos, que se pueden consultar en la Sección 3.3.

O4. Elección de un razonador de reglas para datos semánticos

Se realizará un estudio sobre los razonadores de reglas de inferencia para datos semánticos. Las reglas definidas deberán ser ejecutadas por un razonador que generará los resultados finales del proceso de medición.

O5. Desarrollo de una aplicación de prueba de concepto

Elaborar un pipeline que haga uso de la extensión previamente desarrollada. Para ello se deberá implementar la arquitectura de la aplicación teniendo en cuenta las restricciones de la tecnología a utilizar.

Capítulo 3

Estado del Arte

EN el presente capítulo se describen los conceptos teóricos que son necesarios para establecer las bases de la elaboración del PFC. La descripción incluye conceptos como Calidad de Datos (DQ), Datos Abiertos (OD), Datos Enlazados Abiertos (LOD), Web Semántica y Big Data. También se describe un conjunto de tecnologías que se han utilizado para el desarrollo del proyecto.

Web Semántica

En [BLHL01] se define la Web Semántica como una evolución o extensión de la Web tradicional en la que la información es dada mediante significados bien definidos, lo que facilita el procesamiento automático del contenido y permita a las personas y a los ordenadores trabajar en cooperación.

Este concepto ha ido evolucionando y refinándose con el paso del tiempo. En los siguientes apartados se introduce el concepto de Web Semántica y se explicarán sus principales características.

Concepto

El Consorcio para la World Wide Web (W3C) en [W3Cb] define la Web Semántica como:

Una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida. Al dotar a la Web de más significado y, por lo tanto, de más semántica, se pueden obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común, mediante la cual, es posible compartir, procesar y transferir información de forma sencilla. Esta Web extendida y basada en el significado, se apoya en lenguajes universales que resuelven los problemas ocasionados por una Web carente de semántica en la que, en ocasiones, el acceso a la información se convierte en una tarea difícil y frustrante.

[SBLH06] define a su vez la Web Semántica como sigue:

La Web Semántica es una Web de información procesable - información derivada de los

3. ESTADO DEL ARTE

datos mediante una teoría semántica para la interpretación de símbolos. La teoría semántica proporciona una noción de “significado” en la que la conexión lógica de términos establece la interoperabilidad entre sistemas.

Por otro lado, [HBLM02] definen nuevamente la Web Semántica de la siguiente manera:

La Web Semántica es una extensión de la Web actual, en la que a la información disponible se le otorga un significado bien definido que permita a los ordenadores y a las personas trabajar en cooperación. Está basada en la idea de proporcionar en la Web datos bien definidos y enlazados, permitiendo que aplicaciones heterogéneas localicen, integren, razonen y reutilicen toda la información presente en la Web.

Por lo tanto se puede establecer que la Web Semántica es una extensión de la Web en la que se dota de capacidad de anotar información semántica a los datos de manera que proporcionen un significado. En la última definición aparece el concepto de “datos enlazados” como precursor de lo que posteriormente se considerará Linked Data (LD), con la idea de vincular datos mediante estándares de Web Semántica para alcanzar los siguientes objetivos:

1. Reutilizar entidades ya definidas en el modelo de Web Semántica.
2. Hacer de estas entidades conceptos únicos, evitando redundancia y ambigüedad en los datos.
3. Permitir la interoperabilidad semántica entre aplicaciones heterogéneas.

La reutilización es posible gracias al concepto de Identificador Uniforme de Recurso (URI) utilizado para identificar de forma unívoca, universal y expansible un espacio de nombres de recursos de información.

Respecto de la interoperabilidad semántica, [HT06] explica que la *Web Semántica es la solución al problema de la integración de datos* sin necesidad de llevar a cabo procesos de conversión, gracias a la utilización de un modelo de representación como Framework de Descripción de Recursos (RDF) y utilizando Lenguaje de Marcas Extensible (XML) como fuente sintáctica para su intercambio.

Terminología de Web Semántica

- **Identificador Uniforme de Recurso (URI):** cadena de caracteres que identifica los recursos de una red de forma unívoca. La diferencia con una Localizador Uniforme de Recurso (URL), es que éstos pueden variar en el tiempo. En el ámbito de la Web Semántica, una URI identificará a un recurso de manera unívoca dentro del conjunto de datos.
- **Recurso:** Se dice que un recurso es cualquier concepto que se pueda identificar.
- **Tripla:** Usando el estándar Framework de Descripción de Recursos (RDF), se define

tripla como una asociación de dos recursos a través de una relación o propiedad. Esta asociación se representa mediante dos nodos conectados por un arco (véase figura 3.1) (sujeto, predicado y objeto), también llamado *sentencia* (statement):

- **Sujeto:** es el recurso desde el que parte el arco (la propiedad).
 - **Predicado:** es la propiedad que etiqueta el arco.
 - **Objeto:** es el recurso o literal apuntado por el arco.
- **Endpoint:** Interfaz que se ofrece como extremo de una comunicación o como terminal que permite dar un servicio determinado. En el ámbito de la Web Semántica, un Endpoint permitirá tener acceso a las URI de un dataset, así como a las triplas mediante la utilización de protocolo HTTP.
 - **Graph/Named Graph:** El término *Graph* se refiere a un conjunto de triplas relacionadas entre sí, de tal forma que tanto predicados como objetos dentro de una tripla hacen referencia a sujetos de otras triplas, enlazándose de esta manera. Un *Named Graph* es un conjunto de triplas que tiene sentido en sí misma (por ejemplo, las triplas que definen a un grupo de música en concreto).
 - **Almacenamiento de triplas:** Es una base de datos especializada en almacenar archivos semánticos, es decir, conjuntos de triplas o *graphs*. Su funcionamiento interno dista del modo en que lo hacen las bases de datos convencionales, pues debe permitir inferencia gracias a las propiedades descritas.
 - **Servidor de triplas:** Un servidor de triplas es una aplicación que, dado un almacenamiento de triplas, permite que ese contenido sea accesible por otras aplicaciones y mediante protocolos tales como HTTP. Además debe permitir operaciones de consulta, actualización, inserción y borrado sobre los datos almacenados.

Estándares

A continuación, se expondrán los estándares que definen la Web Semántica en el marco de la W3C.

Lenguaje de Marcas Extensible (XML)

En [XML] se define XML como un formato de texto simple, muy flexible, derivado de Standard Generalized Markup Language (SGML) (ISO 8879). Originalmente diseñado para cumplir con los desafíos de la publicación electrónica a gran escala, XML también está desempeñando un papel cada vez más importante en el intercambio de una amplia variedad de datos en la Web y otros sistemas.

XML establece las bases para la elaboración de lenguajes orientados a la representación de información estructurada mediante la descripción de gramáticas, permitiendo diferentes niveles de abstracción.

3. ESTADO DEL ARTE

La principal ventaja de XML es que permite la intercomunicación de aplicaciones y la integración de información, también en el ámbito de las bases de datos.

XML constituye la base de otros lenguajes y en particular, fue precursor de RDF.

Framework de Descripción de Recursos (RDF)

Se define RDF en [RDF] como un modelo estándar para el intercambio de datos en la Web. RDF tiene características que facilitan la fusión de datos incluso si los esquemas subyacentes difieren y soporta específicamente la evolución de esquemas con el tiempo sin necesidad de realizar cambios en los consumidores de los datos.

RDF permite extender la estructura de los enlaces de la Web para hacer uso de las URI como nombre de las relaciones entre conceptos. De esta manera, se establecen lo que se conoce como *tripas* como una relación (*subjeto*, *predicado*, *objeto*) (véase figura 3.1 y Sección 3.1.2) en la que todos los componentes son URI.

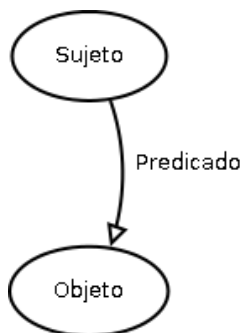


Figura 3.1: Esquema de tripla

Usando este modelo se permite la mezcla de datos estructurados y semi-estructurados a través de diferentes aplicaciones.

Esta estructura de enlaces forma grafos etiquetados y dirigidos donde los arcos representan el tipo de relación entre dos recursos, representados por nodos del grafo (véase figura 3.2).

Lenguaje de Ontología Web (OWL)

Según [OWL], se define OWL como un lenguaje de Web Semántica diseñado para representar conocimiento enriquecido y complejo acerca de conceptos, grupos de conceptos y relaciones entre conceptos. OWL es un lenguaje computacional basado en la lógica de manera que el conocimiento que expresa puede ser explotado por programas de computador.

Los documentos generados según el lenguaje OWL se conocen como **ontologías**. Las ontologías pueden ser publicadas en la W3C o pueden referirse o derivarse de otras ontologías.

En el contexto de la computación y ciencias de la información, una ontología define un

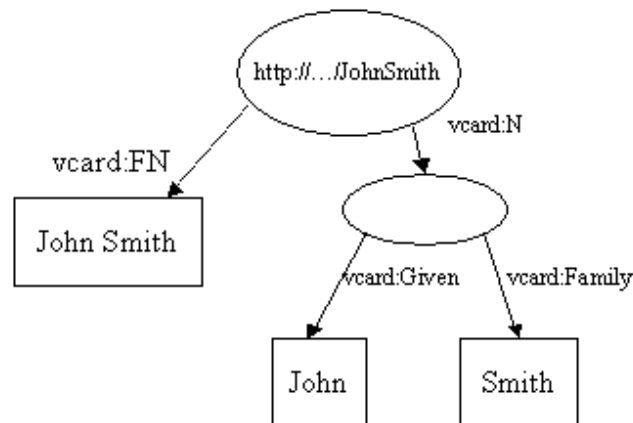


Figura 3.2: Grafo basado en triplas. Extraído de [JEN14]

conjunto de primitivas de representación con la que modelar un dominio de conocimiento [Gru]. La finalidad de las ontologías es ofrecer un modelo formal acerca de un conjunto de conceptos sobre el cual poder aplicar razonamiento automático.

De OWL derivan tres sub-lenguajes basados en la capacidad expresiva, tal y como cita [OWL]:

1. **OWL Lite:** que da soporte a las necesidades básicas del usuario cuando lo que se presente es una representación no tan exhaustiva, como por ejemplo, recursos, clasificaciones jerárquicas y restricciones simples.
2. **OWL DL (Description Logics):** soportando más expresividad semántica y garantiza que todas las inferencias puedan ser calculadas en un tiempo finito.
3. **OWL Full:** el grado de expresividad es total, pero no asegura que las inferencias puedan ser calculadas en un tiempo finito.

SPARQL Protocol and RDF Query Language (SPARQL)

SPARQL es un lenguaje estándar para la consulta de grafos RDF. En [SPAb] se puede encontrar toda la información y especificación de la tecnología.

Muy similar a Structured Query Language (SQL), es un lenguaje declarativo que permite estructurar las consultas como patrones de *triplas* sobre los cuales extraer instancias concretas.

Arquitectura

Tal y como se puede ver en las figuras 3.3 y 3.4, la arquitectura de la Web Semántica se fundamenta sobre los principios de la Web tradicional. Estos principios se pueden resumir:

3. ESTADO DEL ARTE

1. Utilización de Localizador Uniforme de Recurso (URL) para la localización de recursos
2. Uso de HyperText Markup Language (HTML) para la elaboración de documentos, de modo que sea entendible por las personas y procesable por los computadores.
3. Uso de HyperText Transfer Protocol (HTTP) de comunicación cliente-servidor.

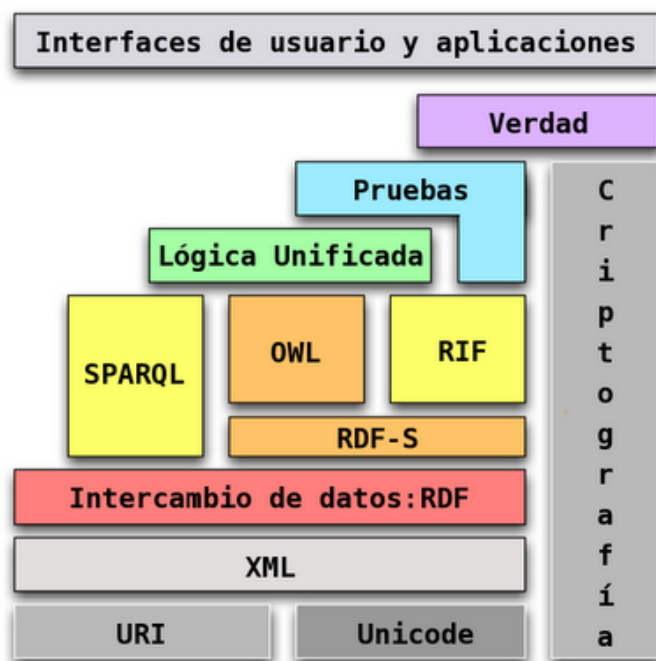


Figura 3.3: Tecnologías de Web Semántica

Las diferentes capas que se muestran en la figura 3.4 [San11] se pueden definir como :

- **Capa de localización y codificación:** el estándar para la codificación de caracteres es UNICODE y para la identificación y localización de recursos se utilizarán las URI o URL.
- **Capa de sintaxis:** estándares necesarios para la representación de información. Cobra especial importancia XML puesto que ofrece un formato de fácil procesamiento con una sintaxis jerárquica. Como derivación de XML surge XML *namespaces* que habilitará la aparición de diferentes vocabularios XML en un mismo documento. Esto permite la reutilización de recursos que previamente ya se hayan definido.
- **Capa de descripción y estructura:** el uso de RDF como estándar de representación de recursos, propiedades y relaciones, es el siguiente gran salto en la Web Semántica. Al tener una representación de la semántica formal, se consigue la interoperabilidad semántica entre aplicaciones heterogéneas.

- **Capa de integración lógica de ontologías y reglas de inferencia:** estrechamente vinculada con la capa de descripción y estructura, en esta capa se amplía la definición de clases, relaciones y propiedades entre recursos, utilizando para ello el lenguaje específico OWL.
- **Capa de consulta:** los recursos ya representados junto con sus relaciones ya conforman un hito. A continuación se requiere establecer unos mecanismos para búsqueda de triplas. SPARQL cumplirá la función de motor de búsqueda declarativo sobre archivos RDF.

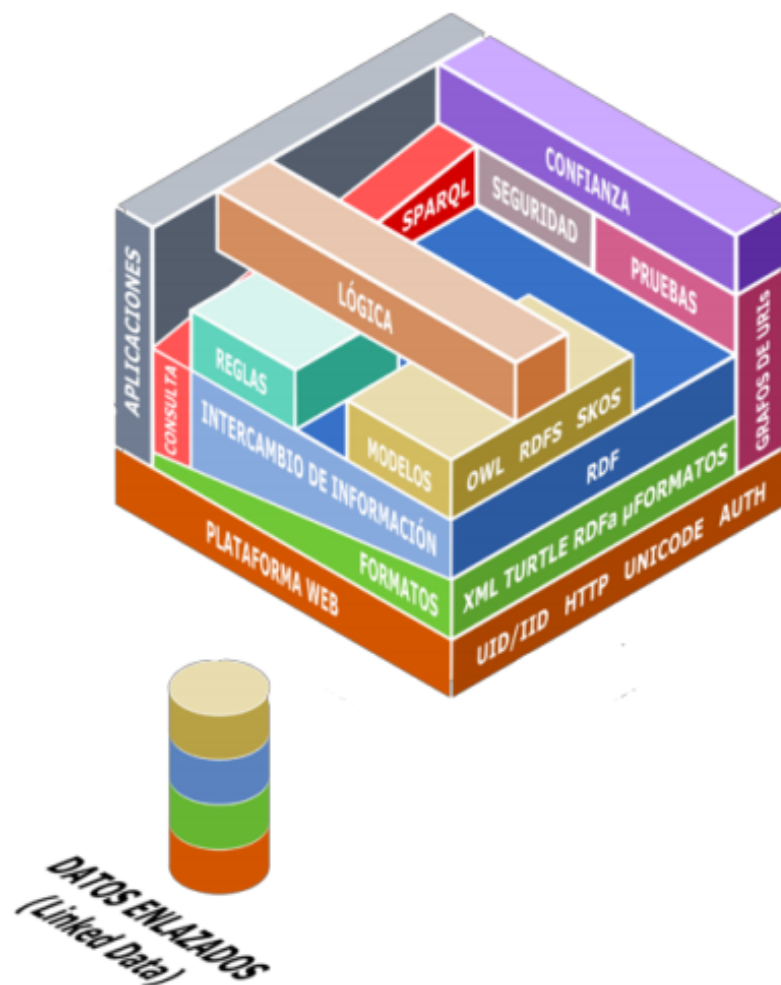


Figura 3.4: Arquitectura de la Web Semántica. Extraído de [San11]

Ontologías

Las ontologías pueden jugar un papel crucial para permitir el procesamiento del conocimiento, el intercambio y la reutilización basada en la Web entre aplicaciones. Las ontologías ofrecen una comprensión común de conceptos, con el fin de que la comunicación entre personas y aplicaciones sea homogénea [DMVH⁺00].

Definiciones de ontologías

[Gru] define las ontologías como *una especificación explícita de una conceptualización*, refiriéndose este término al modelo abstracto de una realidad concreta.

En [SBF98] se profundiza en los términos que engloba la definición de ontología:

Una ontología es una especificación formal y explícita de una conceptualización compartida, donde:

- el término “conceptualización” se refiere a un modelo abstracto de algún fenómeno en el mundo, identificando los conceptos relevantes del mismo,
- “explícita” porque los tipos de conceptos utilizados así como las constantes en su uso están explícitamente definidas,
- “formal” se refiere al hecho de que la ontología debe ser legible para las computadoras, lo que excluye al lenguaje natural, y
- “compartida” refleja la noción de que una ontología captura el conocimiento consensuado, es decir, no es privativo para ningún individuo, sino que es aceptado por un grupo.

Luego una ontología es una jerarquía que define una serie de clases, atributos y relaciones para describir un dominio sobre un concepto en concreto cuya finalidad es servir de herramienta para la representación del conocimiento.

Componentes de una ontología

En [San11] se enumeran los distintos componentes de las ontologías:

- **Clases:** conceptos generales acerca de un determinado dominio. La idea básica que se pretende formalizar.
- **Relaciones:** enlace entre conceptos (clases) de un mismo dominio.
- **Atributos:** representan la estructura del concepto.
 - **Funciones:** identifican un elemento mediante el cálculo de una función.
- **Instancias:** representa un individuo concreto perteneciente a una clase.
- **Axiomas:** expresiones siempre ciertas sobre relaciones.

Clasificación de las ontologías

A su vez, dependiendo del objetivo de la ontología, se pueden clasificar en distintos ámbitos tal y como propone [Gua98] (véase figura 3.5):

- Ontologías de alto nivel (*Top-level ontologies*): describen conceptos muy generales que son independientes de un problema particular, tales como espacio, tiempo, objeto, evento o acción.

- Ontologías de dominio y tarea (*Domain ontologies and task ontologies*): describen respectivamente, el vocabulario relacionado con un dominio genérico o con una tarea genérica o actividad, especializando los términos de la ontología de alto nivel.
- Ontologías de aplicación (*Application ontologies*): describen conceptos dependiendo tanto del dominio particular como de la tarea, es decir, es una especialización que generalmente particulariza *ambos* tipos de ontología. Generalmente corresponde con *roles* desempeñados por entidades de dominio mientras desarrollan alguna tarea.

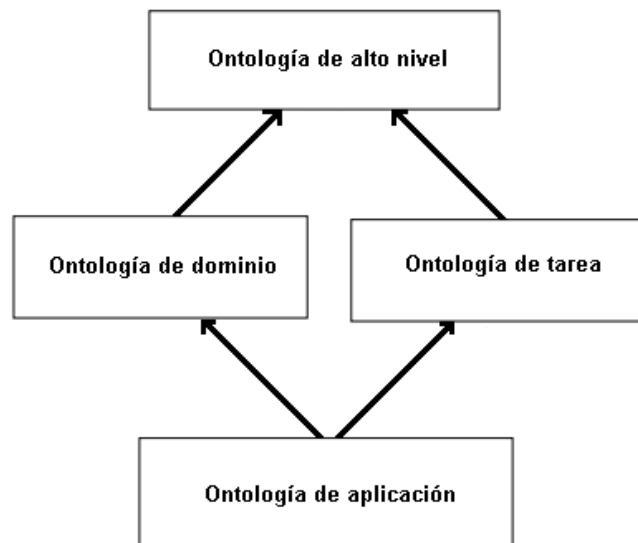


Figura 3.5: Clasificación de Ontologías. Extraído de [Gua98]

Metodología para desarrollo de una ontología en entornos de Web Semántica

Finalmente, [NMo01] propone una secuencia de tareas que todo desarrollo de cualquier ontología debería incluir:

1. Definir las clases en la ontología
2. Organizar dichas clases según una taxonomía
3. Definir propiedades de las clases y los valores que se asocian a esas propiedades
4. Completar los valores de las propiedades para cada una de las instancias.

Vocabularios

Como complemento al concepto de ontología, se describe a continuación el término vocabulario, frecuentemente usado en gran parte de la bibliografía.

Definición

En la Web Semántica, tal y como se explica en [W3Ca], los vocabularios definen los conceptos y relaciones usados para describir y representar un área de conocimiento y así clasificar los términos en una aplicación particular caracterizando relaciones y restricciones. Los vocabularios en la práctica pueden ser enormemente complejos o muy simples (llegando a describir únicamente uno o dos conceptos).

Según esta definición no queda claro en qué se diferencia un vocabulario de una ontología, pues realmente no existe una división clara entre estos dos conceptos. La tendencia es utilizar el término “ontología” para una colección de términos más compleja y formal, mientras que “vocabulario” se usa cuando la flexibilidad en el formalismo no implica necesariamente una pérdida de significado [W3Ca].

Por lo tanto, la función de un vocabulario es similar a la de una ontología en términos de objetivo final: establecer una representación formal de un determinado concepto.

Ejemplos de vocabulario

Existen numerosos vocabularios en uso actualmente. A continuación se citan algunos ejemplos:

Description Of A Project (DOAP)

El objetivo de este vocabulario es la descripción de proyectos software. Para ello, incluye toda la terminología referente al proyecto (licencia, versión de producto, dirección de repositorio, ...).

Simple Knowledge Organization System (SKOS)

Este vocabulario tiene por objetivo la representación y estructuración de esquemas conceptuales tales como taxonomías, esquemas de clasificación o tesauros.

Friend Of A Friend (FOAF)

Quizá uno de los más extendidos sea Friend Of A Friend (FOAF) [FOA]. Su finalidad es describir personas, relaciones entre ellas así como aspectos de su actividad. Esta tecnología está en alza debido a su extensión en las redes sociales. FOAF a día de hoy es la base de un número considerable de esfuerzo en lo que se conoce como movimiento “open social”, que tratan de facilitar al usuario integrar su propia información a través de aplicaciones sociales a través de la Web [AH08].

FOAF trabaja según el principio de *Anyone can say Anything about Any topic* (AAA). En el caso de FOAF los temas usualmente son otras personas [AH08].

En la figura 3.6 y en el listado 3.1 pueden verse ejemplos de uso de este vocabulario. En la figura se aprecian dos recursos (`foaf:ian` y `foaf:mary`), dos propiedades (`foaf:knows` y `foaf:firstName`) y un literal (“Mary”). En el listado 3.1 se expone un documento describiendo algunas entidades y relaciones.

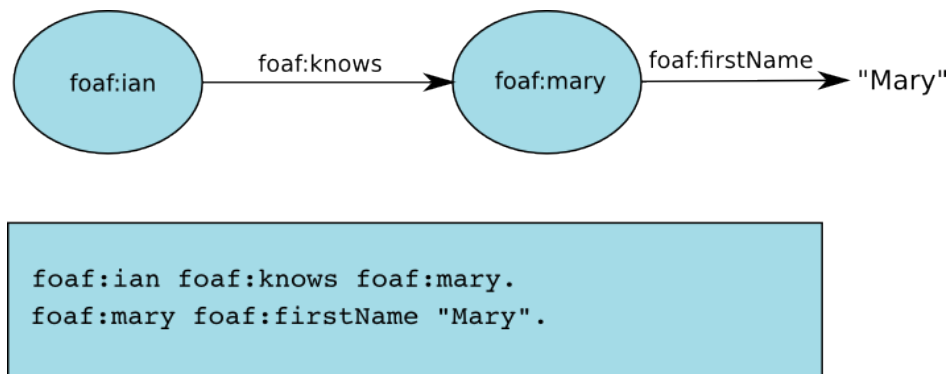


Figura 3.6: Ejemplo de FOAF. Extraído de [JEN14]

```

1 <rdf:RDF
2   xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
3   xmlns:rdfs='http://www.w3.org/2000/01/rdf-schema#'
4   xmlns:foaf='http://xmlns.com/foaf/0.1/'>

6   <foaf:Person>
7     <foaf:name>Raul Reguillo Carmona</foaf:name>
8     <foaf:title>Mr</foaf:title>
9     <foaf:nick>Radulfr</foaf:nick>
10    <foaf:weblog rdf:resource='http://geeklandtryp.blogspot.com.es' />
11    <foaf:knows>
12      <foaf:Person>
13        <foaf:name>Ismael Caballero</foaf:name>
14      </foaf:Person>
15    </foaf:knows>
16  </foaf:Person>
17 </rdf:RDF>
  
```

Listado 3.1: Ejemplo de FOAF

Frameworks para el desarrollo de aplicaciones de Web Semántica

En [W3Cc] se puede consultar una lista con todos los frameworks de Web Semántica disponibles hasta la fecha. En este apartado, se van a nombrar solamente algunos de ellos por ser especialmente significativos.

Apache Jena

Jena es un framework para la construcción de aplicaciones que utilizan tecnologías semánticas y Linked Data [JEN14].

En la arquitectura de Jena (véase figura 3.7) se pueden encontrar tres bloques diferenciados en función de las herramientas que engloban:

■ RDF

- Jena ofrece una API para el tratamiento con grafos RDF, serializando las triplas y tratando con ellas de manera ágil y eficiente.

3. ESTADO DEL ARTE

- Paralelamente ofrece un motor SPARQL para las consultas sobre los archivos semánticos.

■ Triple Store

- Para hacer persistentes los datos, Jena ofrece Triple Data Base (TDB): una base de datos de triplas nativa de alto rendimiento y alineada con el resto de tecnologías de Jena.
- Por otra parte, Jena ofrece **Fuseki** como *endpoint*, es decir, como servidor de triplas accesible a través de HTTP, integrándose a la perfección con TDB.

■ OWL

- Para trabajar con modelos y OWL, Jena propone una API específica orientada a ontologías.
- Jena igualmente propone una API para facilitar el razonamiento y comprobar el contenido de los archivos semánticos. Permite especificar distintos razonadores.

A continuación se van a definir algunos de los conceptos que maneja Jena así como se va a introducir brevemente el ámbito de la inferencia en este framework.

Modelo

A la hora de trabajar con triplas, jena utiliza los llamados **modelos**. Un modelo de Jena es un conjunto de triplas RDF con el que se puede trabajar de diversas formas: bien para hacer consultas sobre él o para aplicar inferencia. Estos modelos son la piedra angular de la tecnología puesto que son la fuente del resto de operaciones que se desarrollan en este framework.

Inferencia

La inferencia es un proceso abstracto de derivación de información adicional partiendo de los datos [JEN14]. De esta manera se utilizará la inferencia para hacer explícita información que aparece de forma implícita en los datos.

La inferencia en Jena consta de dos partes fundamentales:

- **Razonadores:** encargados de llevar a cabo la inferencia, pudiendo encontrar en Jena los siguientes [JEN14]:
 1. Razonador transitivo: ofrece soporte para el razonamiento a través de los conceptos de clase y propiedad. Únicamente afecta a propiedades transitivas y reflexivas de `rdfs:subPropertyOf` y `rdfs:subClassOf`.
 2. Razonador de reglas RDFS: Implementando un subconjunto configurable de vínculos RDFS.
 3. Razonadores para OWL, OWL Mini, OWL Micro: Un conjunto no completo para la implementación de inferencia basada en OWL/Lite y OWL/Full.

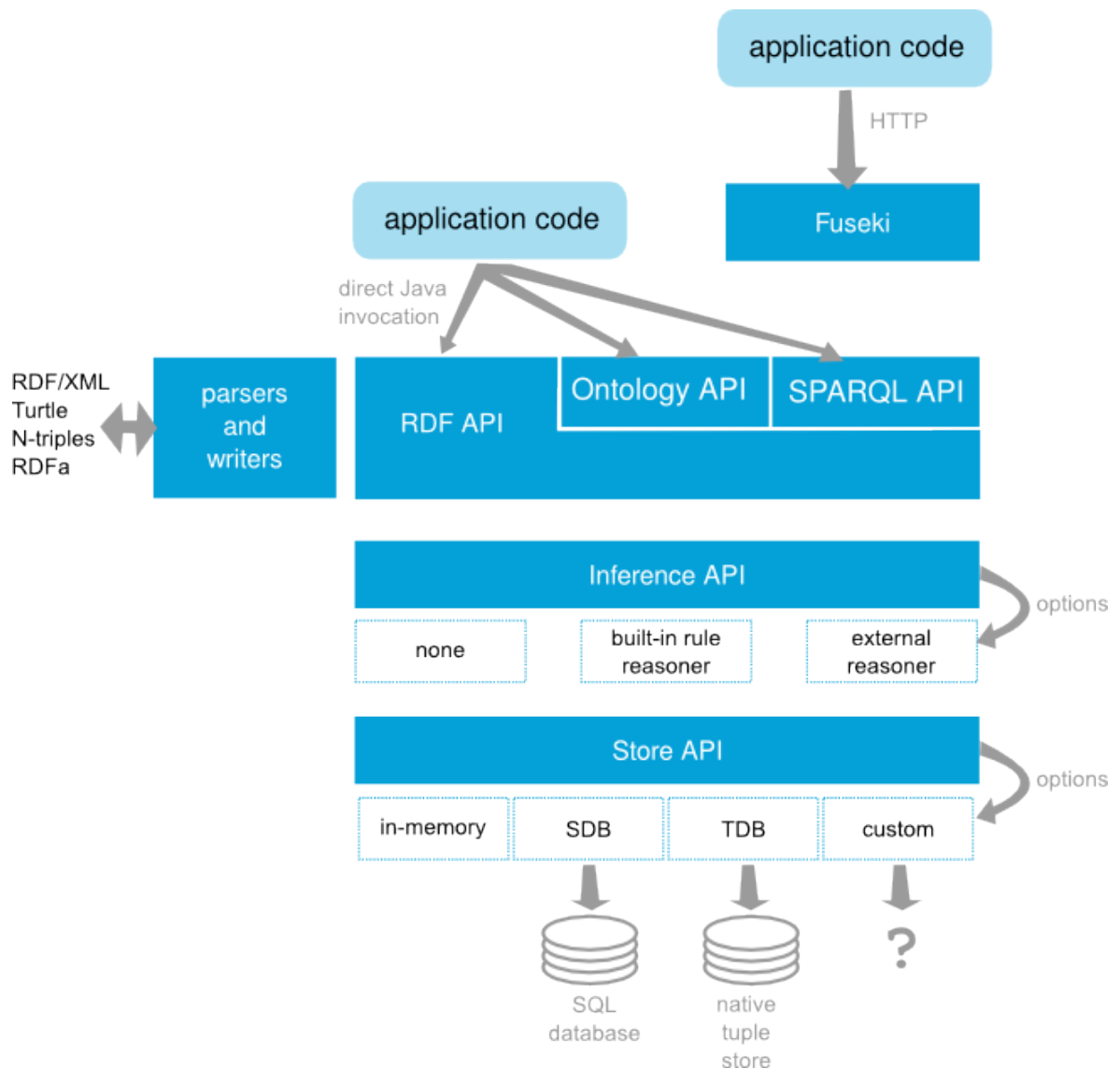


Figura 3.7: Diagrama de la arquitectura de Jena. Extraída de [JEN14]

4. Razonador de reglas genéricas: Se basa en reglas definidas, mediante el en-cadenamiento hacia adelante, hacia atrás e híbrido.

- **Reglas:** las reglas son sentencias usadas para aplicar inferencia en un determinado modelo. Jena implementa un lenguaje propio para la edición de estas reglas.

Se puede consultar un esquema del razonamiento en Jena en la figura 3.8.

Sesame

OpenRDF Sesame es un framework estándar *de-facto* para el procesamiento de datos RDF [OPE]. Sesame incluye operaciones para el procesamiento, consulta, almacenamiento e inferencia sobre RDF. Al igual que Jena, Sesame contiene un número considerable de herramientas para la construcción de aplicaciones de tecnologías semánticas.

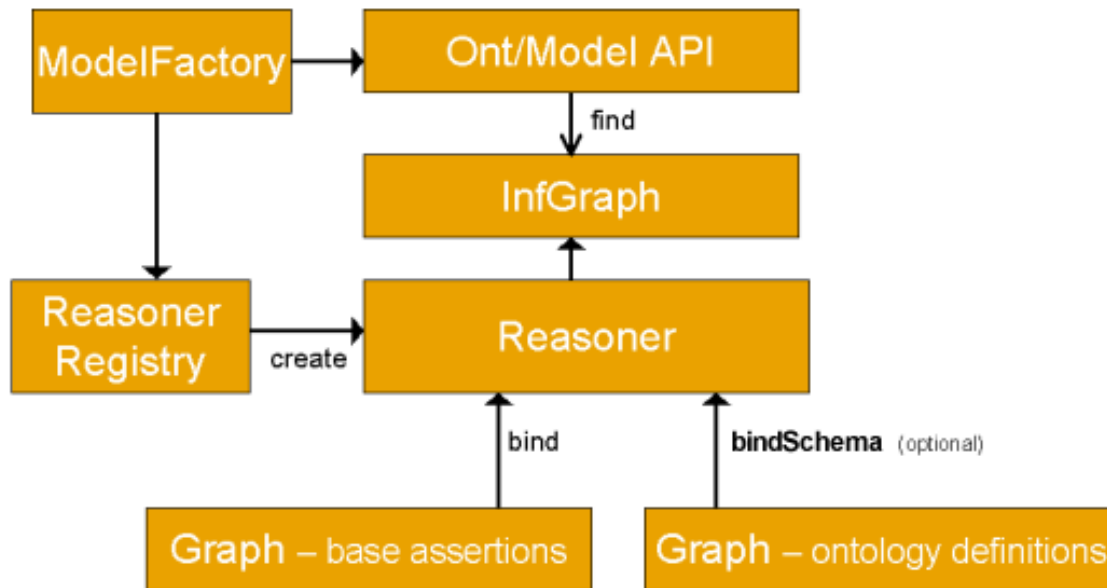


Figura 3.8: Razonamiento en Jena. Extraída de [JEN14]

CubicWeb

CubicWeb [CUB] es otro framework, *open source*, para la realización de aplicaciones semánticas. Está escrito en Python. Sus características engloban:

- Soporta OWL y RDF
- Relationship Query Language (RQL)
- Herramientas de migración para desarrollo ágil
- Una librería de *cubes* como pequeños módulos al estilo de Ruby.

Comparativa

En el cuadro 3.1 se pueden contemplar algunas características de los frameworks anotados anteriormente, pudiendo encontrarse una relación más completa en [W3Cc]:

Framework	Lenguaje	Licencia	Versión
Apache Jena	Java	Apache License 2.0	2.11.0 / September 18, 2013
Sesame	Java	BSD-style license	2.7.11 / March 27, 2014
CubicWeb	Python	Lesser General Public License	3.18.4 / April 7, 2014

Cuadro 3.1: Comparativa de frameworks de Web Semántica. Extraída de [W3Cc]

Datos Enlazados (LD)

En esta sección se incluyen los conceptos clave respecto de los principales movimientos de datos abiertos y datos enlazados.

Definición de LD

Berners-Lee en [BL09] enfatiza la publicación de los datos en la Web no únicamente como exposición de los mismos, sino mediante el establecimiento de enlaces de forma que personas o máquinas puedan explorar una Web de datos. De esta manera se pueden encontrar datos relacionados.

La mayor parte del contenido en la Web está construido con documentos. Estos documentos a su vez tienen enlaces hacia otros documentos cuyo contenido puede estar o no formalizado. El movimiento LD pretende dar un paso más allá, estableciendo relaciones entre los datos de manera global (véase figura 3.9). Para ello, RDF y las URI toman un papel crucial. [BL09] expone sus cuatro reglas para LD:

1. Usar URI como nombres para los conceptos.
2. Usar HTTP para que las personas puedan acceder a esos nombres.
3. Proporcionar información útil cuando la URI sea desreferenciada, usando estándares como RDF o SPARQL.
4. Incluir enlaces a otras URI, de manera que los usuarios puedan descubrir nuevos conceptos.

[BHBL09] establece a su vez una serie de pasos básicos para la publicación de LD:

1. Asignar URI a las entidades descritas por el conjunto de datos y proveer el desreferenciado de las URI a través del protocolo HTTP en representaciones de RDF.
2. Establecer enlaces RDF a otros recursos de datos en la Web, de tal manera que los usuarios puedan navegar a través de la Web de los datos siguiendo enlaces RDF.
3. Facilitar metadatos sobre los datos publicados, de manera que los usuarios puedan evaluar la calidad de los datos publicados y escoger entre diferentes medios de acceso.

Aprovechando LD la información en la Web Semántica puede verse desde diferentes niveles de granularidad, desde el grafo universal formado por todos los documentos RDF en la Web, pasando por documentos individuales hasta sus triplas [DFP⁺05] (véase figura 3.10).

Datos Abiertos (OD) y Datos Enlazados Abiertos (LOD)

Por otra parte los Datos Abiertos (OD), se presentan en ausencia de formatos privativos, información pública y reutilizables procedentes de organizaciones tales como las gubernamentales. Así, Datos Enlazados Abiertos (LOD) serán todos aquellos datos enlazados que

3. ESTADO DEL ARTE

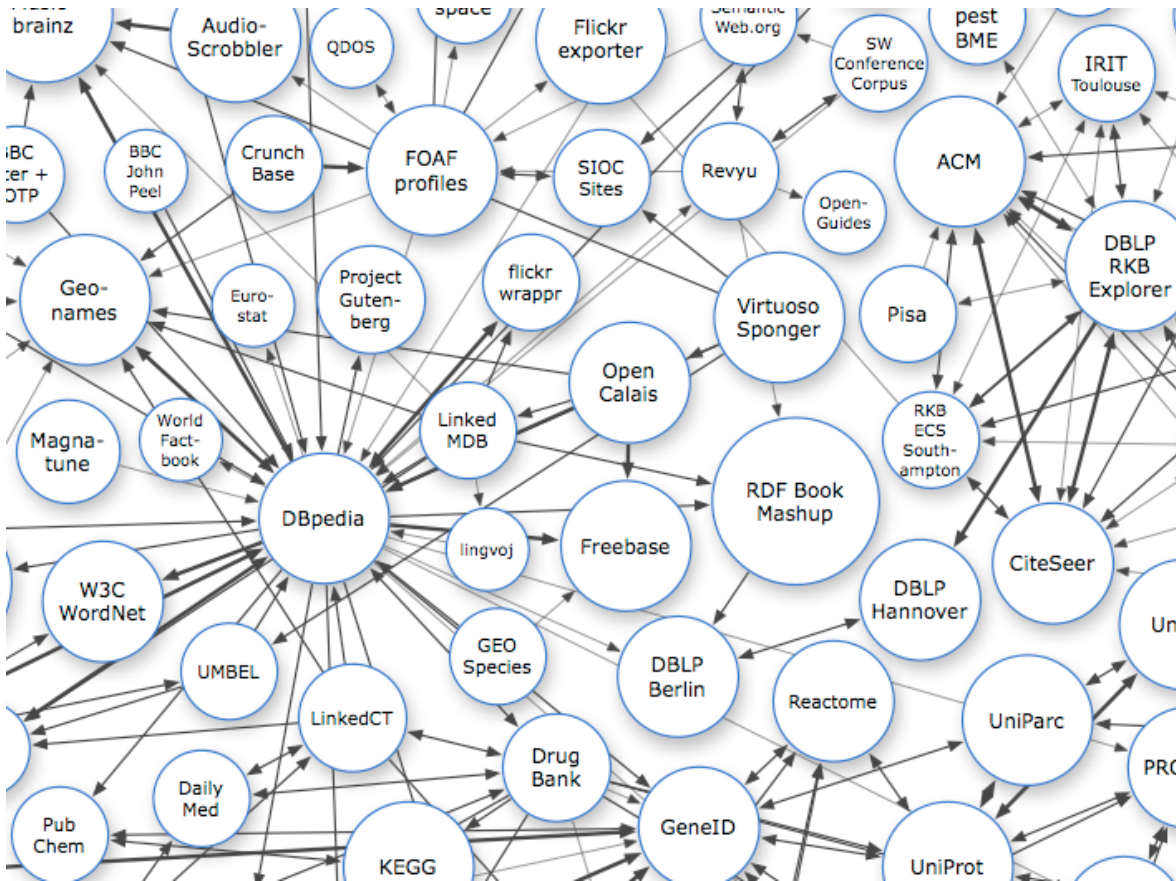


Figura 3.9: Esquema LD DBpedia

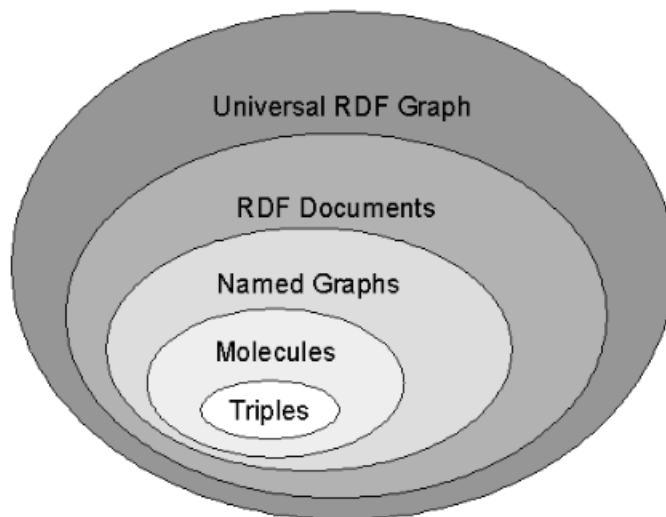


Figura 3.10: Niveles de granularidad en la Web Semántica. Extraído de [DFP⁺05]

sean abiertos. [BL09] define LOD como LD que es liberado bajo licencia abierta que no impida su reuso de manera gratuita.

Asimismo, define las cinco estrellas de calidad de LOD:

- ★ Los datos están disponibles en la Web, independientemente del formato, pero con licencia abierta, para que sean OD.
- ★★ Disponible para la lectura por parte de máquinas, es decir, datos estructurados. Por ejemplo, utilizar una tabla Excel en lugar de una imagen escaneada de dicha tabla.
- ★★★ Como el anterior, pero en formato no propietario. En este caso formato Comma Separated Values (CSV) en lugar de Excel.
- ★★★★ Todo lo anterior y además, usar estándares abiertos de W3C (RDF y SPARQL) para identificar conceptos, de manera que otros usuarios puedan apuntar a este contenido.
- ★★★★★ Lo anterior y además, enlazar estos datos a los de otros usuarios para proporcionar un contexto.

LOD en la actualidad

Empresas del sector privado y público se han dado cuenta de las ventajas que OD y LOD pueden ofrecerles. Es por ello que están surgiendo muchas aplicaciones tanto a nivel nacional, donde estos movimientos aún no están demasiado extendidos, como a nivel internacional. En [APP] se encuentra una lista con una serie de aplicaciones realizadas utilizando tecnología OD y LOD. A continuación se exponen algunos ejemplos.

Bus Guru

Bus Guru es una aplicación para iOS que monitoriza en tiempo real la situación de los autobuses urbanos así como su hora de llegada estimada para una parada (véase figura 3.12).

Bus Gijón

Igualmente en el ámbito nacional, existen aplicaciones que aprovechan los datos abiertos generados en tiempo real por dispositivos empotrados en autobuses y marquesinas. Un caso homólogo a **Bus Gurú** es **Bus Gijón** (figura 3.11), que recaba información del portal de OD <https://datos.gijon.es> para obtener información de los autobuses.

Calidad de Datos

En [SLW97] se identifican tres roles a la hora de tratar los datos:

1. *Data producers*: aquellas entidades encargadas de **generar** los datos.
2. *Data custodians*: personas responsables del **almacenamiento y procesado** de los datos.
3. *Data consumers*: personas o grupos que **usan** los datos.

Debido a que los diferentes roles tendrán concepciones distintas de lo que los datos significan para ellos según su papel en un sistema, pueden encontrarse distintas perspectivas de lo

3. ESTADO DEL ARTE

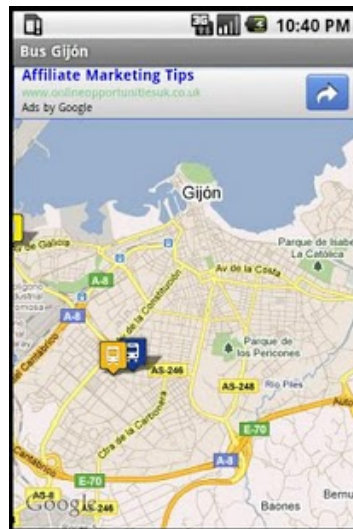


Figura 3.11: Bus Gijón App

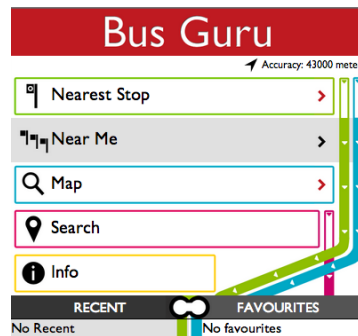


Figura 3.12: Bus Gurú App

que significa calidad de datos. Dos de las principales perspectivas son *Meeting Requirements* y *Fitness for Use*.

- *Meeting requirements* [BH12]: Los datos tienen calidad si satisfacen los requisitos que fueron establecidos. Los **requisitos de datos** que son especificados, por ejemplo, con un modelo Entidad-Interrelación (ER) en el que se subrayan cómo se van a relacionar los datos como conjunto, es decir, un marco arquitectónico para los datos. En [FH10] y [FH11] se pueden consultar aproximaciones a la calidad de datos desde este punto de vista.
- *Fitness for use* [SLW97]: Se dice que los datos tienen calidad si son válidos para el propósito por el que son requeridos, es decir, considerando la calidad de los datos en el **contexto** de uso. En [CVCP08] se adopta esta aproximación.

Dimensiones de Calidad de Datos

La perspectiva *Fitness for Use* se centra en los *Data consumers*, considerando datos de alta calidad aquéllos que son apropiados al usuario final en el contexto de uso. Debido a ello, se deben considerar aspectos tales como *utilidad* o *usabilidad* y en definitiva, todo aquel aspecto que pueda repercutir en la experiencia del usuario.

Puesto que la calidad de los datos dependerá del propósito, se requiere considerar una serie de **categorías**. La calidad de datos es un concepto que es necesario evaluar desde distintos criterios o **dimensiones** (Dimensión de Calidad de Datos (DQD)) [SLW97]. Al conjunto de dimensiones de calidad de datos utilizados para evaluar un conjunto de datos se le conoce como **Modelo de Calidad de Datos**. En el cuadro 3.2 se detallan conjuntos de dimensiones de calidad agrupadas por categorías.

Categoría de Calidad de Datos	Dimensiones de Calidad de Datos
Intrínseca	Precisión, Objetividad, Credibilidad, Reputación
Accesibilidad	Accesibilidad, Acceso seguro
Contextual	Relevancia, Valor añadido, Temporalidad, Completitud, Cantidad de datos adecuada
Representacional	Interpretabilidad, Facilidad de entendimiento, Representación concisa, Representación consistente

Cuadro 3.2: Categorías y dimensiones de DQ. Extraído de [SLW97]

Pese a que existan modelos de calidad de datos consistentes, la calidad de datos no deja de ser un concepto subjetivo, puesto que para un mismo conjunto de datos se pueden obtener niveles de calidad radicalmente distintos dependiendo del usuario o el rol que haga uso de ellos, incluso sobre la misma dimensión de calidad [Wan98].

A continuación se detallarán algunas dimensiones de calidad de datos.

Completeness

Dentro de todas las posibles dimensiones de calidad, existen subconjuntos que cobran especial relevancia. *Completeness* es, por norma general para todos los autores, una dimensión de obligada presencia en todos los trabajos.

Dentro de esta dimensión, en la bibliografía se proponen distintas métricas que contemplan el concepto de *completitud* desde diversas perspectivas. Como se comprobará en adelante (véase Sección ??), en el presente trabajo se ha optado por una única visión, más compacta, sencilla y acorde a lo que un usuario puede esperar acerca de la presencia de valores no nulos en Linked Data.

Definición

[ZRM⁺13] define **Completeness** como el grado en el que toda la información requerida está presente en un conjunto de datos particular. En general, esta definición se puede extender sobre otros factores tales como profundidad de los datos, anchura y alcance para la tarea que se quiera realizar.

[PLW02] amplía esta definición. Define Completeness como el grado en que los datos no han desaparecido y poseen la suficiente amplitud y profundidad para la tarea en cuestión.

Por otro lado, la norma ISO 25012 (véase [ISO]) define Completeness como el grado en el que los datos asociados con una entidad tienen valores para todos los atributos esperados e instancias relacionadas en un contexto de uso específico.

Accessibility

En entornos de Web Semántica y Linked Data tiene especial importancia el hecho de que dichos datos sean de fácil acceso. La palabra *accesibilidad* cuando se habla de datos enlazados cobra un valor mayor que en otras dimensiones. Es preciso que los datos enlazados estén, en efecto, adecuadamente enlazados.

Definición

ISO 25012 (véase [ISO]) define *Accessibility* como el grado en el que los datos puedan ser accedidos en un contexto de uso específico, particularmente por gente que necesite tecnología de apoyo o una configuración especial debido a alguna discapacidad.

Por otra parte [PLW02] define Accessibility como el grado en que los datos están disponibles y son accesibles fácil y rápidamente.

Calidad de Datos en LD

Actualmente, existen autores tales como [ZRM⁺13] o [FH11] que comienzan a aplicar conceptos de LD para algunas dimensiones de calidad. Para ilustrarlo, se muestran algunas métricas propuestas en estos trabajos para las DQD Completeness y Accessibility.

Métricas para Completeness

[ZRM⁺13] propone una serie de métricas para la evaluación de esta dimensión.

1. *Schema Completeness*: grado en el que las clases y propiedades de una ontología están representadas. También conocido como *ontology completeness*.
2. *Property Completeness*: evaluación sobre los valores perdidos de una propiedad específica.
3. *Population Completeness*: siendo el porcentaje de todos los objetos del mundo real de un determinado tipo que están representados en los conjuntos de datos.

4. *Interlinking completeness*: refiriéndose al grado en el que las instancias en el conjunto de datos están interconectadas. Esta métrica consta de especial importancia en Linked Data. No obstante esta métrica concreta puede entenderse (como así se abordará en este trabajo, en la Sección ??) como una particularización de otra dimensión de calidad bien distinta: *Accesibility*.

[FH10] ofrece un conjunto de consultas SPARQL que permiten identificar problemas de integración de calidad de datos. Se pueden comprobar dos ejemplos en los listados 3.2 y 3.3 extraídos de ese mismo trabajo.

```

1 SELECT ?s
2 WHERE { {
3   ?s a <class1> .
4   ?s <prop1> "". }
5 UNION{
6   ?s a <class1> .
7   NOT EXISTS {
8     ?s prop1> ?value}}}

```

Listado 3.2: Consulta SPARQL para identificación de literales perdidos (I)

```

1 SELECT ?s
2 WHERE { {
3   ?s a <class1> .
4   ?s <prop1> <value1>.
5   NOT EXISTS{
6     ?s <prop2> ?value2 .
7   }
8 }UNION{
9   ?s <prop1> <value1> .
10  ?s <prop2> "".
11 }}

```

Listado 3.3: Consulta SPARQL para identificación de literales perdidos (y II)

Métricas para Accessibility

[ZRM⁺13] considera Accessibility como una categoría que a su vez contiene cuatro dimensiones, cada una con sus métricas:

- **Disponibilidad**: grado en el la información está presente y preparada para su uso. Sus métricas propuestas son:
 - *Accessibility of the server*: comprobación sobre el servidor SPARQL ante una consulta.
 - *Accessibility of the SPARQL endpoint*: comprobación sobre el servidor SPARQL ante una consulta.

3. ESTADO DEL ARTE

- *Accessibility of the RDF dumps*: comprobación sobre la recuperación de datos en un contenedor de datos RDF.
 - *Dereferencability issues*: comprobación en el momento que una URI devuelva un error del código de respuesta o enlace roto.
 - *No structured data available*: detección de enlace caído o cuando una URI sin metadatos RDF o sin redirección, devuelva el código de error pertinente.
 - *Misreported content types*: detección de si el contenido es susceptible a ser consumido y si el contenido puede ser accedido.
 - *No dereferenced back-links*: detección de todos los enlaces propios al conjunto de datos: localmente disponibles.
- **Rendimiento**: Eficiencia del sistema vinculado al conjunto de datos de manera que cuanto más eficiente sea la fuente de datos, un sistema puede procesar más eficientemente los datos. Las métricas para esta dimensión son:
- *No usage of slash-URIs*: uso de URIs abreviadas cuando existen grandes cantidades de datos.
 - *Low latency*: si una petición HTTP es contestada en un tiempo medio de un segundo.
 - *High throughput*: número de peticiones HTTP contestadas por segundo.
 - *Scalability of a data source*: detección de si el tiempo en responder una cantidad de diez peticiones, dividido entre diez, no es mayor que el tiempo en responder una petición.
 - *No use of prolix RDF features*: detección del uso de primitivas RDF tales como contenedores y colecciones.
- **Seguridad**: Grado en el que los datos pueden ser restringidos y de esta manera protegidos contra alteraciones ilegales y uso no permitido. Las métricas propuestas son las siguientes:
- *Access to data is secure*: uso para login de credenciales o uso de protocolos específicos.
 - *Data is of proprietary nature*: el propietario de los datos permite el acceso únicamente a ciertos usuarios.
- **Tiempo de respuesta**: Medición del retardo, generalmente en segundos, entre el envío de una consulta por el usuario y la recepción de los resultados. Existe una métrica para esta dimensión:
- *Delay in response time*: retardo entre el tiempo de envío de una petición por el usuario y la recepción de dicha petición por el sistema.

Big Data

Concepto

Desde los comienzos del siglo XXI han tenido lugar una serie de cambios muy significativos en la industria de las tecnologías de la información, tales como el nacimiento del Internet de las Cosas, Computación en la Nube y la aparición de las Redes Sociales.

El desarrollo y explotación de estas tecnologías ha propiciado que la cantidad de los datos que se generan por unidad de tiempo se haya incrementado a una velocidad sin precedentes, trayendo consigo la necesidad de cómputo de tales volúmenes de información heterogénea, en tiempos razonables y de manera escalable.

Big Data es por lo tanto un concepto que surge debido a la necesidad del uso de herramientas y técnicas específicas para procesar esta información, imposible para los sistemas transaccionales comunes [BAR].

De esta manera, el concepto *Big Data* se puede caracterizar según la definición de las tres V's:

- **Volumen:** en referencia al tamaño del conjunto de datos. No hay un número estándar que defina a partir de qué tamaño del conjunto de datos se pueda empezar a hablar de Big Data. No obstante, cuando el conjunto de datos no es susceptible de ser procesado en un único nodo, y requiere la presencia de varios de éstos, puede decirse que es Big Data. Luego el orden de magnitud puede ir desde los cientos de GB hasta los Petabytes.
- **Velocidad:** referenciando a las necesidades relativas al consumo o proceso de datos en tanto se aproximen al tiempo real (*real time*) o próximos al tiempo real (*near real time*).
- **Variedad:** en relación a la heterogeneidad de los datos. Esto incluye representaciones diversas de datos estructurados, como diversas fuentes relacionales, datos semiestructurados o datos no estructurados, como pudieran ser imágenes o vídeo.

Adicionalmente se ha extendido esta definición a medida que las tecnologías y la problemática ha cambiado, encontrando:

- **Valor:** considerando el beneficio que se puede extraer de los datos bajo la premisa que tanto la complejidad de la extracción del valor, como dicho beneficio en sí mismo se incrementa junto con el volumen.
- **Veracidad:** entendiendo como el grado de confianza que se establece sobre los datos.

Cada dimensión deja de manifiesto un problema concreto que hay que afrontar en un contexto Big Data. No obstante estas definiciones, a medida que las tecnologías evolucionan, se adaptan y renuevan, pudiendo hablar sobre la existencia de hasta 9Vs.

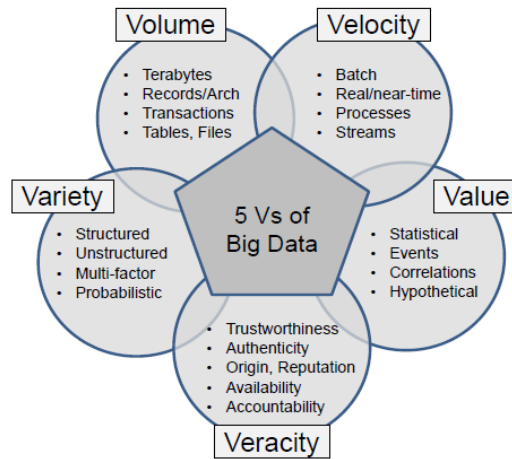


Figura 3.13: 5vs del Big Data

En [DG04] se define por primera vez un modelo de procesamiento y generación de grandes conjuntos de datos, a través de funciones *map* y *reduce*, apoyándose en que muchos problemas del mundo real son expresables en estos términos.

En adelante se desarrollaron multitud de frameworks y herramientas que basándose en estos principios optimizarán y permitirán el procesado escalable de grandes volúmenes de información, cumpliendo con el paradigma de las 5Vs. Entre ellos, se encuentra el ecosistema Hadoop que se describirá con detalle en la sección Ecosistema Apache Hadoop.

Este modelo consiste en el particionado de los datos de entrada y procesado en paralelo de dichas particiones por varios nodos, gestionados por un nodo maestro que ejerce de árbitro y orquestador. En la etapa final, la unificación de los resultados parciales en torno a una solución final se lleva a cabo por otro conjunto de nodos que combinarán el resultado en uno o varios archivos de salida.

Al cierre del 2015, los miembros de la ITU (Unión Internacional de Telecomunicaciones), aprobaron la primera norma de Big Data, estableciendo la primera definición oficial [ITU]:

Paradigma para hacer posible la recopilación, el almacenamiento, la gestión, el análisis y la visualización, potencialmente en condiciones de tiempo real, de grandes conjuntos de datos con características heterogéneas.

Desafíos

Cuando la volumetría y la disparidad de los datos está presente, existe una serie de desafíos que deben afrontarse a la hora de acometer un proyecto [CZ14]:

- **Captura y almacenamiento:** debido al ritmo de crecimiento de los datos así como a su origen disperso, el modo en el que se acceden, capturan y almacenan representa una

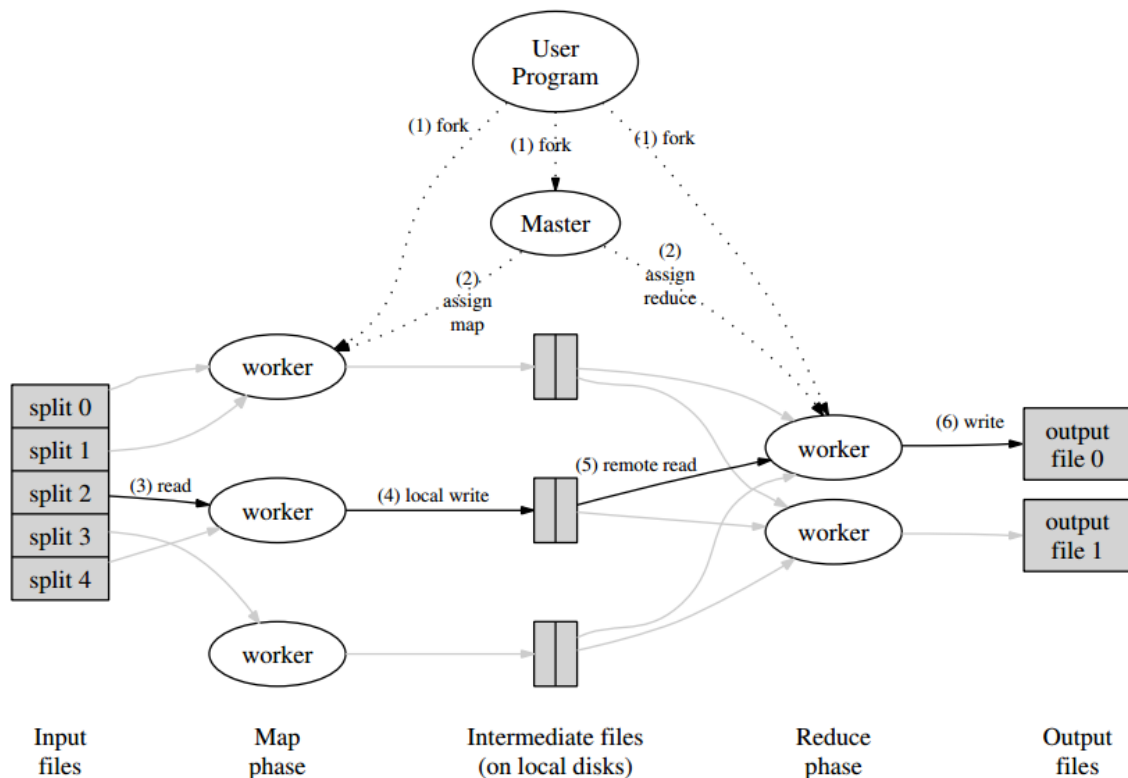


Figura 3.14: MapReduce: ejecución [DG04]

primera barrera a superar.

- **Descubrimiento y limpieza:** referente a tratamiento de la información previo a su procesamiento, análisis exploratorio, gestión de datos perdidos o corruptos.
- **Transmisión:** el ancho de banda es un factor limitante cuando se habla en un contexto de Big Data y puede representar el cuello de botella de un sistema si no se gestiona correctamente el movimiento de tales volúmenes de datos.
- **Análisis y visualización:** llevar a cabo un análisis de volúmenes masivos de datos implica desarrollar técnicas tanto a nivel hardware como software para permitir dichos análisis. Igualmente la visualización resulta un factor crucial a la hora de obtener lecturas fiables de cara a la interpretabilidad de dichos datos.

Frameworks

Existe una gran cantidad de frameworks y ecosistemas dedicados al procesamiento distribuido de grandes volúmenes de información. Entre ellos caben destacar el ecosistema Hadoop y las soluciones de Amazon Web Services, que serán descritos a continuación.

Ecosistema Apache Hadoop

El ecosistema Hadoop se ha convertido en el estándar *de facto* a la hora de llevar a cabo proyectos Big Data. Se organiza a través de una serie de módulos encargados de diversas operaciones en el contexto de un cluster Big Data [HAD].

Entre sus principales componentes caben destacar los siguientes:

- **HDFS:** Hadoop Distributed File System, como sistema de archivos distribuido y escalable.
- **Hadoop MapReduce:** framework para escribir aplicaciones que puedan procesar grandes volúmenes de información en paralelo en el contexto de un cluster, de una manera tolerante a fallos, siguiendo el modelo MapReduce.
- **YARN:** Yet ANother Resource Negotiator es un framework para la gestión de recursos de un cluster, encargado de balancear la carga, asignar nodos a los distintos trabajos y liberar recursos.
- **Pig:** plataforma de alto nivel para crear programas que se ejecutan sobre HDFS, a través de un lenguaje de programación de scripts que permiten realizar operaciones ETL (Extract, Transform and Load) sobre datos guardados en HDFS.
- **Hive:** proyecto que se ejecuta sobre HDFS con el fin de establecer una plataforma de consulta y análisis a través de una interfaz SQL.
- **Oozie:** orquestador de tareas dentro del ecosistema Hadoop.
- **Sqoop:** herramienta para el traspaso eficiente de datos entre HDFS y otras soluciones de almacenamiento, tales como bases de datos relacionales o ficheros de texto plano.
- **Flume:** solución de recolección y distribución de grandes volúmenes de datos entre distintas fuentes.
- **HBase:** base de datos distribuida, escalable y no relacional.
- **Zookeeper:** servicio centralizado para el mantenimiento, la configuración, nombrado y sincronización de los distintos servicios ejecutándose en el contexto de un cluster Hadoop.

Posteriormente se han ido agregando a proyectos apache nuevos frameworks destinados a cubrir necesidades específicas o a mejorar el rendimiento de los actuales:

- **Spark:** [SPAa] es un framework para el procesado de grandes volúmenes de información. La ventaja principal respecto de Hadoop MapReduce es que el procesamiento se realiza en memoria en lugar de en disco, con lo cual la velocidad se incrementa en varios órdenes de magnitud. Ofrece una serie de herramientas de alto nivel:

- **SparkSQL**: módulo para el procesamiento distribuido de datos estructurados. Ofrece una interfaz parecida a consultas SQL y mantiene información estructural acerca de los datos distribuidos.
- **MLlib**: librería de machine learning que incluye una serie de algoritmos distribuidos, supervisados y no supervisados, pipelines, persistencia y utilidades para dar soporte al proceso de aprendizaje automático de manera distribuida.
- **GraphX**: es uno de los componentes más recientes de Spark. Incluye una abstracción *Graph* que trabaja directamente sobre los dataset distribuidos permitiendo operaciones básicas en el manejo de grafos.
- **Spark Streaming**: es una extensión del núcleo de Spark, que introduce el concepto de mini-batch como unidad operativa básica y ofrece un procesamiento distribuido y tolerante a fallos en streaming. Los datos pueden ser ingestados desde distintas fuentes, tales como Kafka, Kinesis o sockets TCP.
- **Kafka**: plataforma de streaming distribuida, ideada para dar soporte a aplicaciones real-time.
- **Mahout**: framework que ofrece una serie de algoritmos distribuidos de Machine Learning.

Soluciones Amazon Web Services

Amazon Web Services es una colección de servicios en la nube, homólogos a los que se encuentran en el ecosistema Hadoop, que establece una plataforma de computación.

Entre los principales caben destacar:

- **S3**: Simple Storage Service, equivalente a HDFS, ofrece un sistema de archivos distribuido, altamente escalable e integrado en los demás servicios.
- **EC2**: Elastic Compute Cloud, ofrece capacidad de cómputo adaptable a través de servidores que pueden realizar autoescalado en los centros de datos de amazon, versátiles para hospedar sistemas software.
- **EMR**: Elastic Map Reduce proporciona un marco de Hadoop hospedado que permite hospedar datos en instancias EC2 escaladas dinámicamente.
- **Kinesis**: equivalente a Kafka, permite la recopilación, procesamiento y el análisis de datos streaming en tiempo real.

Datos Semánticos en entornos Big Data

Si bien los frameworks anteriormente descritos ofrecen un conjunto de herramientas de procesamiento de datos de propósito general, no existe a día de hoy una versión madura de éstos que se haya convertido en un estándar para el procesamiento de grandes volúmenes de datos semánticos. Si bien existen un par de referencias que están comenzando a abrir esta vía:

3. ESTADO DEL ARTE

- **Jena Elephas:** como parte del stack de Jena, Elephas es un conjunto de librerías que proveen de herramientas para el procesamiento de datos RDF sobre Hadoop. No obstante el módulo se encuentra actualmente en una fase Beta [JEN].
- **Sansa Stack:** Scalable Semantic Analytics Stack, es un motor de procesamiento tolerante a fallos que permite computación distribuida sobre datos RDF. Actualmente en la versión 0.2 y en desarrollo [SAN].

Capítulo 4

Método de Trabajo

Para el desarrollo del trabajo, el alumno deberá seguir algún proceso o metodología afín al problema que pretende resolver. Para ello, deberá aportar una pequeña descripción de los procesos o metodologías y justificar su adecuación al problema a resolver.

Se recomienda empezar, antes de hablar de métodos o técnicas específicos, por los aspectos de organización del trabajo (gestión del proyecto, planificación, esfuerzo, calendario, etcétera).

A modo de referencia, este capítulo tendrá una longitud no superior a 15 páginas.

Capítulo 5

Resultados

En esta sección se describirá la aplicación del método de trabajo presentado en el capítulo4 en este TFM concreto, mostrando los resultados (artefactos software y/o hardware, documentos de diversos tipos: modelos, diagramas, especificaciones, diseños, documentos, manuales, etc.) más importantes.

Este capítulo no tiene una extensión recomendada, sino que depende en mayor medida de los contenidos a presentar. No obstante, se recomienda que la longitud total del documento no supere las 120 páginas.

Capítulo 6

Conclusiones

Breve resumen de lo más destacable del trabajo con la solución propuesta. Análisis del logro del objetivo general y objetivos parciales propuestos. Concluir con posibles mejoras, ampliaciones o trabajos relacionados que quedan por hacer y que tienen interés para el tema tratado.

A modo de referencia, este capítulo tendrá una longitud aproximada de entre 2 y 10 páginas.

Conclusiones

Propuestas de trabajos futuros

Publicaciones

Opinión personal

ANEXOS

Anexo A

Ejemplo de anexo

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Referencias

- [AH08] Dean Allemang y James Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
- [APP] Your City Needs These 7 Open Data Apps. url: <http://mashable.com/2012/11/07/open-data-city-apps/>.
- [BAR] ¿Qué es Big Data? url: <https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/>.
- [BH12] Peter Benson y Melissa Hildebrand. *Managing Blind: A Data Quality and Data Governance Vade Mecum*. ECCMA, Bethlehem (Pensylvania), 2012.
- [BHBL09] Christian Bizer, Tom Heath, y Tim Berners-Lee. Linked Data - The Story So Far:. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. url: <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jswis.2009081901>.
- [BL09] Tim Berners-Lee. Linked-data design issues. W3C design issue document, June 2009. <http://www.w3.org/DesignIssue/LinkedData.html>. url: <http://www.w3.org/DesignIssues/LinkedData.html>.
- [BLHL01] Tim Berners-Lee, James Hendler, y Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, Mayo 2001. url: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>.
- [CMC08] Ismael Caballero, M.A. Moraga, y Coral Calero. Integración de Aspectos de Calidad de Datos en Sistemas de Información. 2008.
- [CUB] CubicWeb Semantic Web Framework. url: <http://www.cubicweb.org/>.
- [CVCP08] Ismael Caballero, Eugenio Verbo, Coral Calero, y Mario Piattini. DQRDFS - Towards a Semantic Web Enhanced with Data Quality. En José Cordeiro, Joaquim Filipe, y Slimane Hammoudi, editors, *WEBIST (1)*, páginas 178–183. INSTICC Press, 2008. url: <http://dblp.uni-trier.de/db/conf/webist/webist2008-1.html#CaballeroVCP08>.

- [CZ14] C. L. Philip Chen y Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.*, 275:314–347, 2014. url: <https://doi.org/10.1016/j.ins.2014.01.015>.
- [DFP⁺05] Li Ding, Tim Finin, Yun Peng, Paulo Pinheiro Da Silva, y Deborah L. McGuinness. Tracking rdf graph provenance using rdf molecules. En *Proc. of the 4th International Semantic Web Conference (Poster)*, página 42, 2005. url: <ftp://www.ksl.stanford.edu/local/pub/KSL-Reports/KSL-05-06.pdf>.
- [DG04] Jeffrey Dean y Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. En *Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6, OSDI'04*, Berkeley, CA, USA, 2004. USENIX Association. url: <http://dl.acm.org/citation.cfm?id=1251254.1251264>.
- [DMVH⁺00] Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, y Ian Horrocks. The semantic web: The roles of XML and RDF. *Internet Computing, IEEE*, 4(5):63–73, 2000. url: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=877487.
- [FH10] Christian Fürber y Martin Hepp. Using Semantic Web Resources for Data Quality Management. En Philipp Cimiano y Helena Sofia Pinto, editors, *EKAW*, volume 6317 of *Lecture Notes in Computer Science*, páginas 211–225. Springer, 2010. url: <http://dblp.uni-trier.de/db/conf/ekaw/ekaw2010.html#FurberH10>.
- [FH11] Christian Fürber y Martin Hepp. Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. En *Proceedings of the 1st International Workshop on Linked Web Data Management, LWDM '11*, páginas 1–8, New York, NY, USA, 2011. ACM. url: <http://doi.acm.org/10.1145/1966901.1966903>.
- [FOA] The Friend of a Friend (FOAF) project | FOAF project. url: <http://www.foaf-project.org/>.
- [Gru] Tom Gruber. Ontology (Computer Science) - definition in Encyclopedia of Database Systems. url: <http://tomgruber.org/writing/ontology-definition-2007.htm>.
- [Gua98] Nicola Guarino. *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, volume 46. IOS press, 1998.

- [HAD] Hadoop ecosystem. url: <http://hadoop.apache.org>.
- [HBLM02] James Hendler, Tim Berners-Le, y Eric Miller. Integrating Applications on the Semantic Web, 2002. url: <http://www.w3.org/2002/07/swint>.
- [HT06] Harry Halpin y Henry S. Thompson. One document to bind them: combining XML, web services, and the semantic web. página 679. ACM Press, 2006. url: <http://portal.acm.org/citation.cfm?doid=1135777.1135877>.
- [ISH85] What is Total Quality Control?: The Japanese Way: Amazon.es: Kaoru Ishikawa: Libros en idiomas extranjeros, 1985.
- [ISO] ISO25012. ISO/IEC 25012 x ISO/IEC 25012 Software-Engineering - Qualitätskriterien und Bewertung von Softwareprodukten (SQuaRE) - Modell der Datenqualität. Technical report.
- [ITU] Y.3600:Big data - Requisitos y capacidades basados en la computación en la nube. url: <https://www.itu.int/rec/T-REC-Y.3600-201511-I/es>.
- [JEN] Apache Jena Elephas. url: <https://jena.apache.org/documentation/hadoop/>.
- [JEN14] Apache Jena - Home, 2014. url: <http://jena.apache.org/>.
- [NMo01] Natalya F. Noy, Deborah L. McGuinness, y others. *Ontology development 101: A guide to creating your first ontology*. Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, 2001. url: http://liris.cnrs.fr/~amille/enseignements/Ecole_Centrale/What%20is%20an%20ontology%20and%20why%20we%20need%20it.htm.
- [OPE] openRDF.org: Home. url: <http://www.openrdf.org/>.
- [OWL] OWL - Semantic Web Standards. url: <http://www.w3.org/OWL/>.
- [PLW02] Leo L. Pipino, Yang W. Lee, y Richard Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002. url: <http://dl.acm.org/citation.cfm?id=506010>.
- [RDF] RDF - Semantic Web Standards. url: <http://www.w3.org/RDF/>.
- [SAN] SANS-Stack. url: <http://sansa-stack.net/>.
- [San11] Juan Antonio Pastor Sanchez. *Tecnologías de Web Semántica*, 2011.

- [SBF98] Rudi Studer, V.Richard Benjamins, y Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2):161–197, Marzo 1998. url: <http://linkinghub.elsevier.com/retrieve/pii/S0169023X97000566>.
- [SBLH06] Nigel Shadbolt, Tim Berners-Lee, y Wendy Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101, Mayo 2006. url: <http://dx.doi.org/10.1109/MIS.2006.62>.
- [SLW97] Diane M. Strong, Yang W. Lee, y Richard Y. Wang. Data Quality in Context. *Commun. ACM*, 40(5):103–110, Mayo 1997. url: <http://doi.acm.org/10.1145/253769.253804>.
- [SPAa] Apache Spark. url: <http://spark.apache.org>.
- [SPAb] SPARQL Query Language for RDF. url: <http://www.w3.org/TR/rdf-sparql-query/>.
- [W3Ca] Semantic Web - W3C. url: <http://www.w3.org/standards/semanticweb/>.
- [W3Cb] Semantic Web - W3C. url: <http://www.w3.es/Divulgacion/GuiasBreves/WebSemantica>.
- [W3Cc] Tools - Semantic Web Standards. url: <http://www.w3.org/2001/sw/wiki/Tools>.
- [Wan98] Richard Y. Wang. A Product Perspective on Total Data Quality Management. *Commun. ACM*, 41(2):58–65, Febrero 1998. url: <http://doi.acm.org/10.1145/269012.269022>.
- [WS96] Richard Y. Wang y Diane M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, Marzo 1996. url: <http://dl.acm.org/citation.cfm?id=1189570.1189572>.
- [XML] Extensible Markup Language (XML). url: <http://www.w3.org/XML/>.
- [ZRM⁺13] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Soren Auer, y Pascal Hitzler. Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*, 2013. url: <http://semantic-web-journal.net/system/files/swj414.pdf>.

Este documento fue editado y tipografiado con \LaTeX empleando la clase **esi-tfm** (versión 0.20170913) que se puede encontrar en:
https://bitbucket.org/arco_group/esi-tfg

[respeta esta atribución al autor]

