Semantic (Big) Data Analysis: An Extensive Literature Review

H. Guedea-Noriega, and F. García-Sánchez

***Abstract*—For many years, companies have exploited the data registered in their everyday operations by their transactional systems to obtain useful information and assist in decision-making. To this end, different data analysis techniques and business intelligence strategies have been applied. In recent years, the increase in the volume of data, along with variety in data and the velocity at which data is being produced, has led to the conception of novel processing mechanisms capable of dealing with such huge amount of data, namely, Big Data. The main difficulties associated with Big Data management are linked to its collection and storage, search, sharing, analysis and visualization. The formal underpinnings of Semantic Web technologies enable the automated processing of data through sophisticated inference and reasoning techniques. Semantic technologies have been successfully applied in a number of scenarios for the integration of heterogeneous data, data analysis at the knowledge level, and visualization of Linked Data. In the last few years, a large number of published research papers have explored the benefits in using semantic technologies in data analysis and Big Data. In this paper, we provide a systematic review of the literature in this research area, highlighting the main benefits obtained by the integration of semantic technologies in data analysis and the most challenging aspects that remain to be addressed.**

***Index Terms*—Big Data Analysis, Semantic Big Data Analysis, Systematic Review.**

I. Introducción

El Análisis de Datos (DA, del inglés *Data Analysis*) reúne procesos de inspección, limpieza, transformación y modelado de datos con el objetivo de presentar información útil que sugiera conclusiones y apoye la toma de decisiones [1]. Es imprescindible la integración de datos previo al análisis, que a su vez está estrechamente vinculado con la visualización y la difusión de datos. El DA se utiliza actualmente en numerosas áreas de aplicación como, por ejemplo, negocios [2], ciencias sociales [3], o educación [4], entre otros. Su desarrollo ha impulsado el nacimiento de técnicas y estrategias específicas de análisis de datos entre las que destacan la minería de datos, enfocada en predecir y descubrir nuevo conocimiento [5], y la inteligencia de negocio (BI) [6], centrada en la gestión empresarial y la agregación de valor con la interpretación de datos para la toma de decisiones.

H. H. Guedea-Noriega, Escuela Internacional de Doctorado, Universidad de Murcia, Murcia (Spain), hector.guedea@um.es.

F. García-Sánchez, Departamento de Informática y Sistemas, Universidad de Murcia, Murcia (Spain), frgarcia@um.es.

Generalmente, la arquitectura tecnológica de una solución de análisis de datos está formada por tres componentes principales [7]: (i) herramientas para llevar a cabo los procesos de extracción, transformación, y carga de datos, conocidas como ETL (*Extract, Transform, and Load*), (ii) un repositorio en el que se almacenan los datos que han sido integrados, y (iii) aplicaciones de explotación y visualización de datos, entre las que se encuentran los sistemas de generación de informes, las técnicas de análisis multidimensional (OLAP), y las herramientas de minería de datos que aplican técnicas estadísticas, de inteligencia artificial y simbólicas. El proceso típico de análisis de datos se puede dividir en tres etapas [8], a saber, preprocesado de datos (adquisición, organización y almacenamiento de datos), procesado de datos (aplicación de técnicas de análisis de datos) y visualización de datos (presentación de los resultados del análisis). Estos procesos conllevan algunos retos y dificultades como [6]: (i) la heterogeneidad de las fuentes, atribuido a la calidad variable, para su integración, limpieza y estandarización del ETL, lo que simboliza inconvenientes en implementar DA en tiempo real, (ii) la imposibilidad de integrar datos no estructurados o semiestructurados en repositorios con sistemas de gestión de bases de datos relacionales, y (iii) la necesidad de servidores OLAP para la exposición multidimensional de los datos.

A medida de la creciente disponibilidad y la importancia de los datos provenientes de fuentes heterogéneas, principalmente de Internet (sitios Web, redes sociales, portales multimedia), los procesos tradicionales de análisis de datos se enfrentan a grandes desafíos casi imposibles de solucionar con sus implementaciones, evidenciando una necesidad de consolidar una arquitectura que soporte volúmenes de datos mucho mayores que los que comúnmente manejan los sistemas relacionales [9]. De esta manera surge una nueva estrategia y colección de tecnologías definida como datos masivos (*Big Data*). Los expertos suelen asociar el término *Big Data* a tres características fundamentales de los datos con los que es necesario tratar, a saber, volumen (cantidad en bytes), velocidad (velocidad de creación y utilización, constantemente en tiempo real), y variedad (diferentes tipos de fuentes y datos); son las conocidas 3 Vs del *Big Data* [10]. En los últimos años se han agregado nuevas "V" ligadas a propiedades que se pueden asociar al *Big Data* como son valor, veracidad y visualización; o incluso volatilidad, validez y viabilidad [11]. Gartner define *Big Data* como "*activos de información caracterizados por su volumen elevado, velocidad elevada y alta variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y la toma de decisiones en las organizaciones*" [12].

Existen distintas versiones de arquitectura para los sistemas de análisis de Big Data. Algunas se asemejan al modelo tradicional "adquirir-organizar-analizar-decidir", mientras que otras contienen más información sobre interdependencias, módulos e interacción, o están enfocadas a un producto específico como HP Reference Architecture MapR M5 u Oracle Big Data Architecture [13]. Sin embargo, todas concuerdan y comparten técnicas y características comunes de escalabilidad, elasticidad y alta disponibilidad. En este trabajo, para el proceso de análisis de Big Data se considerarán las fases de adquisición, organización, análisis y visualización/decisión, estrechamente vinculadas a algunas de las 'Vs' del *Big Data* mencionadas anteriormente [14]. Adquisición es la acción de captura de datos, referida al gran ***volumen*** de datos estructurados, semi-estructurados o no estructurados, utilizando sistemas de gestión de bases de datos, procesamiento de transacciones en línea (OLTP), bases de datos NoSQL, archivos HDFS (sistema de archivos distribuido Hadoop[1]) y archivos de todo tipo. La fase de organización está ligada a actividades de categorización y estructuración de la información, referida a la ***variedad***, en donde los datos se extraen, limpian, filtran, cargan y mezclan (datos cuantitativos frente a cualitativos) en los almacenes de datos mediante herramientas ETL, organizados debidamente en tiempo real con la utilización, por ejemplo, del sistema de procesamiento MapReduce[2] de Hadoop. La ***volatilidad*** toma importancia en esta fase por la necesidad de establecer reglas para la disponibilidad de los datos, así como garantizar una rápida recuperación de la información. Análisis se asocia con la capacidad de reacción del proceso de análisis de datos (empleando técnicas como análisis de texto, estadísticas, minería de datos, análisis de gráficos, análisis espacial, etc.) en el menor tiempo posible para presentar los resultados para su visualización, referida a la propiedad de ***velocidad*** de los datos. Por último, visualización/decisión permite evaluar la precisión y calidad de datos para la toma de decisiones, a través de la propia ***visualización*** de relaciones y patrones resultantes, junto con la ***veracidad*** y la ***validez*** de los datos, que permitirá generar ***valor*** a partir de este nuevo conocimiento.

El *Big Data* trae consigo una gran cantidad de retos por su característica que lo hace notable: la avalancha de datos. En palabras de Mayer-Schönberger y Cukier [15]: "*Cada día creamos 2,5 quintillones de bytes de datos, de forma que el 90% de los datos del mundo actual se han creado en los últimos dos años. Nuestro mundo se ha 'datificado'.*"

En este contexto, no existen decisiones específicas sobre qué datos mantener y qué datos rechazar, o cómo almacenar lo que mantenemos confiablemente con los metadatos correctos. La transformación de contenido no estructurado en formato que pueda ser enlazado con otros datos para su análisis es un reto importante. Análisis de datos, organización, recuperación y modelización son otros desafíos fundamentales [16].

La Web semántica [17] ofrece medios para hacer frente a algunos de los desafíos a los que se enfrentan las actuales soluciones de análisis de datos y *Big Data*. RDF (*Resource Description Framework*), que es el lenguaje de modelado de datos del que se sustenta la Web semántica, está basado en grafos y permite modelar datos en forma de tripletas sujeto-predicado-objeto [18]. De esta manera, es posible representar e interconectar tripletas de RDF para construir *linked data*. Para dotar de mayor semántica formal a RDF surgió OWL (*Web Ontology Language*)[19], que permite publicar y compartir datos usando ontologías [20] en la Web, facilitando en su totalidad la interpretación e interoperabilidad del contenido. Las tecnologías semánticas pueden ser vistas como apoyo al descubrimiento, integración, representación y gestión de conocimiento [21]. *Linked Open Data* (LOD) tiene como principal objetivo agregar una capa semántica sobre los datos que sea comprensible por las máquinas para permitir que las computadoras asuman algunas de las tareas de análisis de datos exclusivas, hasta ahora, para los seres humanos [22]. Con RDF, LOD intenta agrupar y transformar el actual panorama de almacenamiento en un grafo de datos interconectado sobre el cual se pueden construir complejas aplicaciones analíticas [23].

La utilización de LOD y las tecnologías semánticas dentro del análisis de datos y el procesamiento del *Big Data* ha traído consigo una serie de desafíos importantes como [9]: (i) extender el almacenamiento actual para incluir una capa semántica, (ii) distribuir operaciones de datos semánticos, (iii) automatizar el análisis de datos semánticos, (iv) los archivos RDF como formatos de serialización con alto nivel de detalle son particularmente difíciles de almacenar, intercambiar o consultar, por lo cual se atribuyen retos de *Big Semantic Data* (volumen y variedad).

Esta revisión sistemática de literatura tiene como objetivo estudiar las fases de los procesos de análisis de datos que más pueden beneficiarse del empleo de las tecnologías y soluciones ligadas a la Web semántica, enumerando los avances que se han producido hasta el momento en esta dirección y los retos que quedan por superar. El resto del documento se estructura como se indica a continuación. En la sección 2 se describe la metodología seguida para realizar la revisión sistemática de literatura, incluyendo las etapas de alcance, búsqueda de literatura, y selección. Los resultados del estudio se presentan en la sección 3. Éstos han sido divididos en dos grandes apartados, uno centrado en las aportaciones de la semántica al análisis de datos tradicional, y el otro dedicado al análisis de *Big Data* y la contribución de las tecnologías semánticas en el mismo. Finalmente, en la sección 4 se enumeran las principales conclusiones extraídas de esta revisión sistemática.

II. Método

Una revisión sistemática de literatura tiene como objetivo presentar un análisis objetivo y exhaustivo sobre un tema de investigación utilizando una metodología confiable, precisa y verificable [24]. Para ello se evalúan e interpretan las investigaciones relevantes que se han llevado a cabo acerca de la temática de interés. Las etapas por las que hay que pasar para la realización de la revisión sistemática de literatura son las siguientes [25]:

[1] http://hadoop.apache.org/

[2] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

1) Alcance
Establecer las preguntas de investigación.
2) Planificación
Dividir las preguntas de investigación en conceptos individuales para hacer términos de búsquedas y formular criterios preliminares de inclusión y exclusión, para revisarlos en etapas iniciales de la investigación y durante el proceso de filtrado.
3) Identificación (búsqueda)
Usar los términos de búsqueda para realizar el proceso de búsqueda en, por lo menos, dos bases de datos e inspeccionar cuidadosamente los resultados de la búsqueda.
4) Selección
Crear procesos de categorización y clasificación, exclusión e inclusión.
5) Elegibilidad
Analizar en partes la versión completa de los artículos potencialmente elegibles y extraer la información pertinente para ser incluida.

En este trabajo abordaremos las tareas de identificación, evaluación crítica e integración de los resultados para todos los estudios relevantes y de alta calidad vinculados al área de análisis de datos, *Big Data*, y la aplicación de las tecnologías semánticas en ese contexto. En los siguientes apartados se describen en detalle los pasos seguidos para realizar esta revisión sistemática de literatura.

A. Alcance

Esta revisión sistemática de literatura pretende abordar lo relativo a las preguntas de investigación que se presentan en la Tabla I.

TABLA I
PREGUNTAS DE INVESTIGACIÓN Y SUS FUNDAMENTOS

ID	Preguntas	Fundamentos
PI1	¿Es posible realizar análisis semántico de datos (masivos)?	Determinar la viabilidad de integración de una capa semántica en el análisis de datos (masivos).
PI2	¿En qué etapas del análisis de datos (masivos) tiene sentido aplicar tecnologías semánticas?	Identificar las fases del proceso de análisis de datos (masivos) que se pueden beneficiar del uso de tecnología semántica.
PI3	¿Cuáles son las principales ventajas de aplicar tecnologías semánticas en el análisis de datos (masivos)?	Resaltar los beneficios identificados en el uso de soluciones semánticas en las distintas etapas del proceso de análisis de datos (masivos).
PI4	¿Qué retos plantea el uso de tecnologías semánticas en el análisis de datos (masivos)?	Identificar los principales desafíos ligados al análisis semántico de datos (masivos).
PI5	¿Cuál es el nivel de madurez de las soluciones que emplean tecnología semántica en el análisis de datos (masivos)?	Determinar la madurez y la confiabilidad de la investigación publicada respecto al tema en debate.

B. Búsqueda de Literatura: Planificación e Identificación

Con la utilización del programa *Publish or Perish*[3] de Harzing sobre Google Académico se logró identificar literatura con alto grado de impacto académico combinando calidad y cantidad a través de las citas, descargas, valoración y la métrica de Índice H, la cual nos ayuda a identificar publicaciones con mayor aceptación científica. El Índice H es un sistema propuesto por el físico Jorge Hirsch para la evaluación de la calidad de la producción de un investigador en función de la cantidad de citas que han recibido sus artículos científicos [26]: "*Un científico tiene Índice H si ha publicado h trabajos con al menos h citas cada uno. Por lo tanto, si h de sus Np trabajos recibe al menos h citas cada uno, y los otros (Np - h) trabajos tienen como máximo h citas cada uno.*"

El Índice H que se emplea en el programa *Publish or Perish* es una adaptación del definido por Hirsch, asociado a científicos, para su aplicación sobre los resultados de búsquedas de artículos a partir de palabras clave. En particular, *Publish or Perish* obtiene este Índice H ordenando (en orden descendente) las publicaciones por el número de citas recibidas, enumerándolas para identificar el punto en el que el número de orden coincide con el número de citas recibidas por una publicación. Este número constituye el Índice H. Por ejemplo, un Índice H de 10 significa que 10 de los artículos obtenidos como resultado de una búsqueda han recibido al menos 10 citas cada uno.

TABLA II
BÚSQUEDAS EN PUBLISH OR PERISH

Término	Criterio	Número de Registros	Índice H
semantic big data analysis, big data analysis	Con todas las palabras, entre los años 2010 a 2017	999	177
semantic data analysis	Frase exacta en títulos solamente, entre los años 2010 a 2017	13	2
semantic data analysis	Frase exacta en todo el documento	199	18
semantic big data analysis	Frase exacta en todo el documento	6	1
semantic big data analysis	Con todas las palabras	680	120
big linked data analysis	Frase exacta en todo el documento, entre los años 2010 a 2017	1	1
linked data analysis	Frase exacta en títulos solamente, en todos los años.	16	5
linked data analysis	Frase exacta en todo el documento, entre los años 2010 a 2017	174	16
semantic business intelligence	Frase exacta en títulos solamente, entre los años 2010 a 2017	12	2
ontology big data analysis	Frase exacta en títulos solamente, entre los años 2001 a 2017	6	2
ontology data analysis	Frase exacta en títulos solamente, entre los años 2001 a 2017	158	18
análisis de datos semántico	Frase exacta en todo el documento, entre los años 2010 a 2017	3	0
Total		**2277**	**363**

[3] https://harzing.com/resources/publish-or-perish

En la Tabla II se muestran los términos de búsqueda utilizados para el acceso a la literatura, sus diferentes variantes y los criterios considerados en el proceso. Se obtuvieron un total de 2277 artículos relacionados con las temáticas de búsqueda; 363 de estos artículos están incluidos en el Índice H, que es un número de publicaciones suficiente para la realización de la reseña de literatura. Sin embargo, para la selección asertiva es necesario eliminar duplicados y especificar criterios de exclusión e inclusión tal y como se describe a continuación.

C. Selección de Literatura

Con las etapas del diagrama de flujo PRISMA (ver Fig. 1), se identificaron 2277 estudios relacionados con los términos de búsqueda; sin embargo, se localizaron 505 duplicados, y se excluyeron 1752 por no cumplir los criterios de inclusión que se describen en la Tabla III. Como resultado se obtuvieron 20 publicaciones para la realización de este estudio. Para poder hacer un análisis más detallado, se dividieron los artículos seleccionados en dos grandes grupos ligados a las áreas temáticas de (i) análisis semántico de *Big Data* (ver Tabla IV), y (ii) análisis semántico de datos (ver Tabla V).

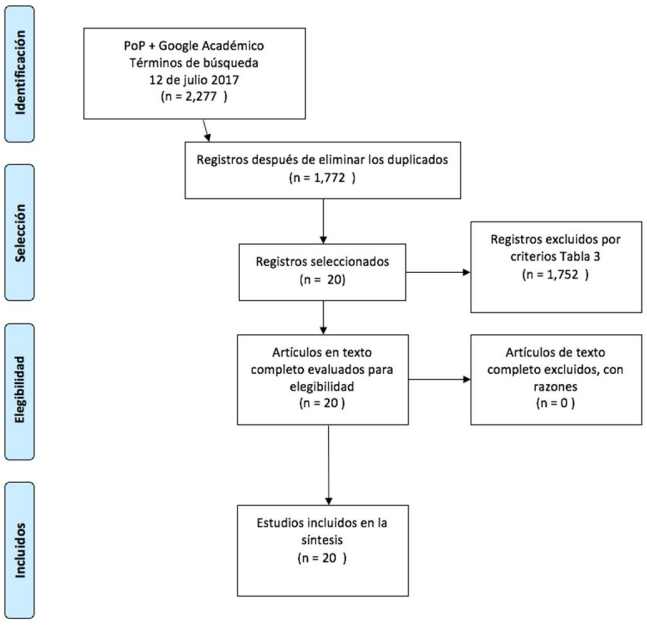

Fig. 1. Diagrama de flujo PRISMA, inclusión y exclusión.

TABLA III
CRITERIOS DE INCLUSIÓN/EXCLUSIÓN

Inclusión	Exclusión
Valoración Índice H.	No contiene métrica Índice H y tampoco es temática relacionada al análisis semántico de datos.
Acceso a resumen y documento completo.	Enlaces rotos a publicaciones.
Literatura en multiformato (libros, revistas, conferencias) publicadas en editoriales de calidad.	Páginas Web apócrifas y de fuente terciaria.
Temática directamente relacionada a Análisis de datos y Semántica.	Ambigüedades en la relación temática.

TABLA IV
ANÁLISIS SEMÁNTICO DE BIG DATA

Nombre de publicación	Año
Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis [27]	2017
Development and Evaluation of an Obesity Ontology for Social Big Data Analysis [3]	2017
Exploration and visualization in the web of big linked data: A survey of the state of the art [28]	2016
Research on ontology modeling of steel manufacturing process based on big data analysis [29]	2016
Social big data: Recent achievements and new challenges [4]	2016
Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions [30]	2015
Semantic link network-based model for organizing multimedia big data [31]	2014

TABLA V
ANÁLISIS SEMÁNTICO DE DATOS

Nombre de publicación	Año
Analysis on Twitter data of automobile domain using ontology [32]	2016
A Framework for Analysis of Ontology-Based Data Access [33]	2016
Mapping Analysis in Ontology-Based Data Access: Algorithms and Complexity Ubiquitous data accessing method in IoT-based information system for emergency medical services [34]	2014
Ontology-Driven Business Intelligence for Comparative Data Analysis [35]	2014
Discovery and visual analysis of linked data for humans [36]	2014
Payola: Collaborative linked data analysis and visualization framework [37]	2013
Formal linked data visualization model [38]	2013
Ontology-based semantic similarity: A new feature-based approach [39]	2012
Aemoo: Exploratory search based on knowledge patterns over the semantic web [40]	2012
Semantic business intelligence-a new generation of business intelligence [2]	2012
Extracting core knowledge from linked data [41]	2011
Ontology-based meta-analysis of global collections of high-throughput public data [42]	2010
Initial Implementation of a Comparative Data Analysis Ontology [43]	2009

D. Elegibilidad: Extracción de Datos y Clasificación

Se extrajeron los datos relevantes sobre las publicaciones de los diferentes grupos temáticos utilizando el formulario de extracción de información que se presenta en la Tabla VI para ser almacenado y ordenado en fichas bibliográficas, con el propósito de sintetizar la información respecto a los objetivos planteados en esta investigación y las preguntas de investigación mostradas en la Tabla I. La información recabada se centra en el modo en que las distintas investigaciones han aportado soluciones a los diferentes desafíos vinculados con la aplicación de las tecnologías semánticas en el análisis de datos (masivos), arquitectura técnica propuesta, metodología seguida y etapa de madurez, así como la viabilidad de la aproximación sugerida.

III. RESULTADOS

En esta sección se presentan los resultados de la revisión sistemática de la literatura sobre los artículos seleccionados con la ayuda del formulario de extracción de información (ver Tabla

VI), donde se describen los diferentes procesos asociados al análisis de datos tradicional y *Big Data* con la utilización de tecnologías semánticas.

TABLA VI
FORMULARIO DE EXTRACCIÓN DE INFORMACIÓN

Elemento	Descripción
Fecha de revisión	La fecha de extracción de datos
Referencias bibliográficas	Autor, año de publicación, título y fuente de publicación
Enfoque o categoría del estudio	El área temática principal o el problema que trata de abordar
Objetivo	Descripción de los objetivos del estudio
Motivación	Explicación de la motivación detrás del estudio
Problema	Problemas que se abordan en el estudio
Solución	Soluciones presentadas en el estudio para resolver problemas planteados
Análisis de datos (masivos)	¿La infraestructura que presenta está basada en análisis de datos tradicional o masivo?
Aplicabilidad semántica en el análisis de datos (masivos)	En qué etapas del análisis de datos (masivo) se emplea tecnología semántica
Beneficios obtenidos de aplicar tecnología semántica	Los beneficios obtenidos como resultado de aplicar tecnología semántica en procesos de análisis de datos
Etapa de madurez	Especificación de la etapa de madurez de la investigación teniendo en cuenta los siguientes puntos: investigación básica, prototipo y herramienta funcional
Limitaciones y futuras direcciones de investigación	Limitaciones y áreas de mejora en el uso de la semántica en el análisis de datos (masivo)

A. Tecnologías Semánticas en el Análisis de Datos

En este apartado se describe el impacto que han tenido las tecnologías semánticas en las etapas principales del proceso de análisis de datos: preprocesado de datos (adquisición, organización y almacenamiento de datos), procesado de datos (aplicación de técnicas de análisis de datos) y visualización de datos (presentación de los resultados del análisis). En cada caso, se enumerarán las ventajas alcanzadas y los retos que quedan por afrontar.

1) Preprocesado

En los últimos años, la Web semántica ha tomado gran importancia y relevancia en aspectos relacionados con la recuperación de la información y adquisición de datos procedentes de diferentes fuentes. El uso de ontologías, núcleo tecnológico de la Web semántica, facilita la extracción de conocimiento significativo y su integración. Las soluciones basadas en ontologías siguen en aumento, con diferentes métodos, enfoques y herramientas que apoyan la extracción de conocimiento. Por ejemplo, se han propuesto sistemas para la construcción automática de ontologías a partir de diferentes tipos de fuentes de entrada (datos estructurados, semiestructurados y no estructurados), sistemas de respuesta a preguntas, extracción de información basada en ontologías, acceso a datos basados en ontologías (OBDA) y diferentes técnicas de minería [33]. Las dificultades de adquirir y almacenar datos heterogéneos tienen solución a través de los sistemas OBDA, los cuales convierten todo tipo de base de datos o fuentes de datos (por ejemplo, las base de datos relacionales, documentos de texto, etcétera) en bases de datos basadas en ontologías (repositorios RDF y OWL), con el objetivo de estandarizar una vista de datos orientada al usuario y hacerla accesible mediante consultas formuladas en un único lenguaje, lo que favorece de forma significativa al acceso futuro de datos.

En los sistemas de BI tradicionales, el papel de los repositorios de datos es organizar, integrar y almacenar datos de todos los departamentos de una organización. Dentro de BI semántico su lugar es ocupado por el repositorio de tripletas. Las tripletas se emplean para realizar afirmaciones sobre recursos Web en el lenguaje RDF [2]. En un escenario de estas características, la base de datos está especialmente diseñada para el almacenamiento y recuperación de tripletas por medio de consultas semánticas. La principal ventaja del uso de RDF en esta etapa del análisis de datos es la posibilidad de obtener información de fuentes de datos heterogéneas, empleando modelos ontológicos para conseguir la integración y soportando fuentes estructuradas, no estructuradas y semi-estructuradas.

En resumen, la principal utilidad que se le ha dado al empleo de tecnologías ligadas a la Web semántica en el contexto del preprocesado de datos es la posibilidad de integrar datos provenientes de fuentes dispares y heterogéneas, ofreciendo un punto de acceso único a los mismos. La organización de los datos en forma de grafos RDF, generando *linked data*, facilita del mismo modo el acceso a datos interrelacionados, independientemente de la fuente de los mismos. Existen, sin embargo, desafíos aparentes como el tiempo de respuesta a consultas, mejorar la eficiencia en los procesos de traducción y esquema de extracción entre las diversas bases de datos para su unificación y, por último, detección de inconsistencias en el entorno del preprocesado de datos para su evaluación y control de calidad.

2) Procesado

Las herramientas de BI existentes son adecuadas para informar y realizar tareas de análisis complejas, pero carecen de una formalización explícita del conocimiento sobre los términos comerciales, la comparación y los procesos de análisis. Una propuesta de mejora es el análisis de datos sobre el repositorio de tripletas realizado a través de la técnica de análisis formal de conceptos [35], donde los datos se estructuran en unidades en forma de abstracciones formales de conceptos del pensamiento humano, permitiendo una interpretación comprensible y facilitando la representación del conocimiento, así como la gestión de la información [2]. En este contexto, es importante generar una capa de ontología multidimensional para consolidar el análisis en diferentes niveles sobre el repositorio de datos, con el fin de agregar un mayor significado a los datos, eliminar redundancias e información irrelevante para obtener un análisis aumentado y sobresaliente [35]. El modelo ontológico, que proporciona una definición inequívoca de los términos del dominio para las necesidades específicas de OLAP, fortalece la correlación y enriquecimiento de los datos de forma automática con algoritmos de agrupamiento, para realizar tareas de análisis

comparativo de datos altamente heterogéneos de diferentes fuentes, plataformas y tecnologías [42]. Las dimensiones semánticas permiten la integración de ontologías de dominio existentes en OLAP. Un repositorio basado en ontologías emplea ontologías para automatizar tareas relacionadas con la construcción de almacenes de datos y procesos de ETL. Con todo, las ontologías sirven como base para el desarrollo de la nueva generación de sistemas de inteligencia empresarial [35].

Existe un gran número de modelos ontológicos expresados en OWL con el objetivo de proporcionar conceptos clave en el análisis comparativo evolutivo, como la CDAO (*Comparative Data Analysis Ontology*) [43]. CDAO está diseñada para facilitar la interoperabilidad de datos e, indirectamente, para permitir el uso más amplio de métodos evolutivos. Este enfoque comparativo se usa comúnmente en bioinformática y otras áreas de la biología para extraer inferencias a partir de una comparación de versiones y transformaciones (por ejemplo, la evolución de una proteína dentro de un organismo cambiante).

El análisis de datos semánticos se puede potenciar mediante búsquedas exploratorias basadas en *Encyclopedic Knowledge Patterns* (EKP), que explotan las técnicas de la Web semántica y la estructura de los enlaces Web para enriquecer los resultados de las consultas con los conocimientos pertinentes provenientes de diversas fuentes (integración y mezcla de información) [40]. Sin embargo, el uso del conocimiento explícito de conjuntos de datos se vuelve insuficiente en sistemas donde solamente utilizan la visualización de *linked data* organizados de forma uniforme. Como solución se pueden emplear *Knowldege Patterns* (KP) conectados para representar conjuntos de *linked data* de forma que sea posible identificar sus componentes básicos de conocimiento a pesar de su orientación uniforme [41].

Por otro lado, los enfoques basados en conocimiento ontológico taxonómico impulsan diversas tareas del procesamiento del lenguaje natural en sistemas informáticos de análisis de datos como la desambiguación del sentido de la palabra, la detección de sinónimos o la detección y corrección automáticas de errores de deletreo, basados en la evaluación del parecido semántico de las palabras [39]. Las aplicaciones directas se pueden encontrar en el campo de la gestión del conocimiento, como la generación de tesauros (lista de palabras o términos para representar conceptos), la extracción de información o el aprendizaje automático de ontologías.

En resumen, utilizar tecnologías semánticas en el análisis de datos favorece la comprensión y representación del resultado como conocimiento, esto simboliza que el proceso de brindar significado a los datos fortalece la eliminación de redundancias e información irrelevante, lo que a su vez desarrolla una ventaja competitiva en la gestión de la información. Sus mayores desafíos son precisamente en el almacenamiento de las tripletas, el diseño de las ontologías adecuadas y en el uso correcto de las técnicas de análisis.

3) Visualización

Hay una gran cantidad de herramientas de visualización de *linked data*, como Tabulator[4], que generalmente muestran datos vinculados como un grafo y también admiten la vista de tabla de las tripletas RDF. Sgvizler[5] permite a los desarrolladores Web presentar resultados de consultas SPARQL como grafos, mapas, etcétera. Las anteriores herramientas se centran únicamente en la parte de representación de la información, ofreciendo un número limitado de técnicas de visualización sin soporte para la extensibilidad. Por otro lado, ViziQuer[6] permite analizar *SPARQL endpoints* desconocidos y obtener una idea acerca de qué datos hay dentro. Ninguna de las herramientas tiene como objetivo brindar soporte a usuarios no expertos y ninguna de ellas proporciona capacidades de análisis y visualización de forma suficiente. En cambio, Payola [37] es una herramienta que integra características de análisis y diferentes formas de visualización, ofreciendo componentes software específicos en forma de *plugins* para ser usados de un modo sencillo por parte de los usuarios, sin necesidad de crear código adicional. Además, brinda la posibilidad de desarrollar nuevos componentes con funciones diversas dentro de la herramienta. La herramienta ha sido desarrollada con Scala, JavaScript y HTML5, lenguajes de programación disponibles en código abierto y relativamente sencillos de aprender.

Los KP antes mencionados también favorecen la visualización y organización del conjunto de datos sobre *linked data* [41]. Las complejidades impuestas por el uso tecnologías semánticas y por el formato de *linked data* dificultan al usuario común la exploración y visualización del conocimiento en dichos datos vinculados. Para solucionar esta problemática, se proponen dos interfaces y/o herramientas basadas en Web dentro de un mismo flujo de trabajo llamado CODE [36], la cual contiene el *QueryWizard* y el *VisualisationWizard* (*VisWizard*). *QueryWizard* hace que buscar en *linked data* sea tan simple como con los motores de búsqueda Web estándar, y proporciona una interfaz tabular que admite transformaciones en el conjunto de datos recuperados (por ejemplo, seleccionar/eliminar columnas, filtrado y agregación). *VisWizard* deriva automáticamente visualizaciones de los conjuntos de datos creados y admite su análisis interactivo utilizando múltiples visualizaciones coordinadas. La ventaja de estas propuestas es que permiten ocultar las complejidades tecnológicas subyacentes a los usuarios, y automatizar el proceso analítico y la extracción de los datos. En general, la utilización de las tecnologías semánticas favorece el análisis interactivo.

Aplicaciones como Aemoo[7] [40] [44] ayudan a filtrar el conocimiento recuperado para mostrar un conjunto razonable y relevante para el usuario, incluida la motivación de porqué se incluye una determinada información. Aemoo utiliza DBpedia[8] como punto de partida para encontrar coincidencias de datos y resolver una consulta del usuario, pero luego procede a enriquecer los datos de DBpedia con información adicional proveniente de otras fuentes como Wikipedia (en particular, su

[4] https://www.w3.org/2005/ajar/tab
[5] http://dev.data2000.no/sgvizler/
[6] http://viziquer.lumii.lv/
[7] http://wit.istc.cnr.it/aemoo/
[8] http://wiki.dbpedia.org/

estructura de enlaces), noticias de Google y Twitter. Aemoo realiza una resolución de identidad en dos situaciones principales: (i) identifica la identidad del recurso al que hace referencia una consulta del usuario, y (ii) determina la identidad de los recursos que se mencionan en las noticias y los *tweets* junto con el tema de la exploración. El uso de EKP permite trazar límites significativos alrededor de los datos. De esta forma, Aemoo realiza tanto el enriquecimiento como el filtrado de información, basándose en la estructura del EKP, que a su vez refleja la forma más común de describir entidades de este tipo particular. El usuario se beneficiará en la forma de presentación por medio de una mejor guía en la navegación de la información: en lugar de navegar un grafo simple de tripletas, navega unidades de conocimiento y se mueve de una a otra sin perder la visión general de una entidad.

En resumen, existen propuestas adecuadas para la exploración, interacción y visualización de la información obtenida durante el análisis de datos semánticos, exponiéndose en diferentes tipos, formatos y estilos, con el firme propósito de crear un recurso de búsqueda y extracción de conocimiento útil para el usuario. Sin embargo, los mayores desafíos en la visualización de datos semánticos están orientados en brindar soporte a usuarios no expertos, con poco o nulo conocimiento técnico de las tecnologías semánticas. La falta de usabilidad de muchas de estas herramientas vuelve más complejo su uso, por tal motivo, se convierten en exclusivas de investigadores o ingenieros de software.

B. Tecnologías Semánticas en el Análisis de Big Data

En este apartado se describe el impacto que han tenido las tecnologías semánticas en las etapas principales del proceso de análisis de *Big Data*: adquisición (captura de datos estructurados, semiestructurados o no estructurados, a través de diferentes sistemas y técnicas), organización (actividades de categorización y estructuración de la información), análisis (análisis de datos con técnicas como análisis de texto, estadísticas, minería de datos, análisis de gráficos, análisis espacial, entre otras, para presentar los resultados para su visualización) y visualización (evaluar la precisión y calidad de datos en la visualización). En cada caso, se enumerarán las ventajas alcanzadas y los retos que quedan por afrontar.

1) Adquisición

Los modelos ontológicos ofrecen un amplio espectro de aplicación en el contexto del *Big Data*. Como se analizó en la sección III.A.1, el acceso a datos basado en ontologías se considera un ingrediente clave para la nueva generación de sistemas de información, especialmente para aplicaciones de la Web Semántica que involucran grandes cantidades de datos. El objetivo es limitar o reducir las consultas sobre bases de datos relacionales. Uno de los usos más interesantes de las conceptualizaciones compartidas es el acceso a datos basados en ontologías (OBDA) [33]. OBDA proporciona una forma conveniente de manejar grandes cantidades de datos distribuidos en fuentes de datos heterogéneas por la formulación de consultas en un único lenguaje. Estas consultas se desarrollan y ejecutan de forma contextualizada en repositorios de tripletas. Las soluciones OBDA en *Big Data* proporcionan un esquema de origen y asignaciones (o mapeo), lo que habilita numerosas oportunidades para conectar la ontología con bases de datos de diferentes tipos y dimensiones. Sin embargo, el mapeo es la parte más compleja de una especificación OBDA, ya que se debe capturar la semántica de las fuentes de datos y expresar dicha semántica en términos de la ontología. Por este motivo se diseñaron algoritmos para establecer límites de complejidad ajustados para los problemas de decisión asociados con la inconsistencia y la redundancia del mapeo durante la asignación del OBDA [34]. Con todo, RDF y OWL se utilizan para crear las ontologías que permitan integrar todos los formatos de datos, y se emplea SPARQL para realizar las consultas que llevan a cabo la extracción, recuperación y manipulación de datos semánticos. Además, existen más lenguajes expresivos como, por ejemplo, RIF (*Rule Interchange Format*) y SWRL (*Semantic Web Rule Language*). En la actualidad existen numerosas herramientas para el desarrollo de sistemas OBDA como DIG Interface[9], Ontop[10], QToolKit[11], MASTRO[12], y un complemento OBDA para Protégé[13] [33].

Algunos autores proponen el uso de metadatos, tanto asociados a archivos como a los contenidos de los mismos, para facilitar el acceso semántico a datos masivos y heterogéneos [45]. Para ello sugieren almacenar los metadatos extraídos empleando el lenguaje OWL en un sistema de archivos distribuido HDFS. En particular, parten de los metadatos representados en XML y aplican transformaciones haciendo uso de mapeos mediante el estándar XSLT. Con esto consiguen salvar las limitaciones de XML como formato de representación (este lenguaje solo considera el ámbito sintáctico) y permiten un acceso más sofisticado a las fuentes de datos sacando provecho de esta representación semántica.

En resumen, existen esfuerzos similares en la extracción de *Big Data* y datos tradicionales con la implementación de soluciones OBDA. La principal diferencia es que en el caso de tratar con *Big Data* es preciso considerar el uso de sistemas de archivos distribuidos para ayudar en la partición y replicación de datos, aportando mayor escalabilidad, tolerancia a fallos y alta concurrencia. En este contexto, también se deben atender nuevos retos en su implementación como son la gestión de operaciones en tiempo real, seguridad, alta disponibilidad y el aumento de la demanda de datos [46].

2) Organización

Las ontologías constituyen el núcleo tecnológico de la Web semántica. Una ontología se puede definir como "*una especificación formal y explícita de una conceptualización compartida*" [20]. El uso de ontologías para modelar los datos del domino y la terminología relevante aporta mayor capacidad potencial de representación y gestión de grandes volúmenes de datos sociales no estructurados relacionados con el dominio [3].

[9] http://dl.kr.org/dig/
[10] http://ontop.inf.unibz.it/
[11] https://github.com/earthlab/qtoolkit
[12] http://www.dis.uniroma1.it/~mastro/
[13] https://protege.stanford.edu/

Esto servirá, además, para mejorar los diferentes tipos de análisis de datos como análisis de sentimientos, minería de datos, minería de textos y más.

Un ejemplo claro del empleo de ontologías para mejorar el modo en que se estructura y organiza la información, permitiendo la integración de datos y su posterior análisis conjunto, es OSHCO (*Oral-Systemic Health Cross-domain Ontology*) [30]. OSHCO es una ontología que modela conocimiento tanto del dominio médico como de la salud bucal y permite realizar complejos procesos de razonamiento e inferencia sobre datos en estos dominios. De hecho, los autores en [30] describen un modelo general sobre cómo la ontologías y los sistemas basados en reglas semánticas pueden aportar valor en el análisis de *Big Data*. Este modelo está dividido en tres capas, a saber, la capa de datos, la capa de conocimiento y la capa de aplicación. En la capa de conocimiento es en la que se situarían la ontologías, reglas y procesos de razonamiento semántico que facilitan el acceso a los *Big Data* de la capa de datos, la cual incluye a su vez una amplia variedad de datos heterogéneos y complejos provenientes de fuentes estructuradas, semi-estructuradas y no estructuradas. Tal y como sugieren para los dominios médico y de salud bucal, gracias al uso de las ontologías, las reglas y los razonadores, será posible realizar inferencias y generar nuevo conocimiento que puede ser empleado para ayudar, ya dentro de la capa de aplicación, en la toma de decisiones, el descubrimiento de servicios y la integración de datos.

Por otro lado, también se ha planteado el uso de tecnología semántica para la organización de grandes cantidades de recursos multimedia [31]. Para ello se propone un nuevo modelo para organizar recursos multimedia con etiquetas sociales denominado *Semantic Link Network* (SLN). La generación de la SLN se realiza mediante el cálculo de la relación relativa entre los distintos recursos multimedia a partir de la relación semántica entre las etiquetas que acompañan a dichos recursos, obtenida empleando diversas métricas. En este caso, la semántica de etiquetas se extrae a partir de búsquedas Web, lo que permite al sistema adaptarse a los nuevos usos de las distintas expresiones empleadas para etiquetar a los recursos. Finalmente, las relaciones explícitas en la SLN se explotan a nivel de aplicación para mejorar la precisión de búsquedas, realizar tareas de agrupación y ofrecer recomendaciones de recursos multimedia relacionados.

Como se puede apreciar en los ejemplos descritos, hacer uso de modelos semánticos para representar los datos facilita la organización de la información estableciendo relaciones explícitas entre los distintos elementos. Esto permite a su vez integrar datos de distintas fuentes y dominios, y realizar inferencias que dan lugar a nuevo conocimiento. Finalmente, a nivel de aplicación la información así organizada se puede explotar de múltiples formas para facilitar la toma de decisiones, la búsqueda de datos interrelacionados, y la integración y agrupación de los datos. Los desafíos en esta etapa del análisis de *Big Data* mediante tecnologías semánticas pasan por definir técnicas y sistemas que permitan llevar a cabo los procesos de razonamiento necesarios sobre grandes cantidades de datos que faciliten la toma de decisiones en tiempo real.

3) Análisis

Se ha demostrado que el análisis de *Big Data* tiene implicaciones para la gestión del conocimiento [47]. Por lo tanto, uno de los desafíos principales es el de mejorar la capacidad de abstracción de información y la transformación a conocimiento para su interpretación. Las tecnologías semánticas ayudan al análisis de *Big Data* en la gestión, intercambio y reutilización de conocimiento. El uso del modelado ontológico con BDAKMS (*Big Data Knowledge Management System*) [29] junto con el establecimiento de reglas SWRL brindará soporte a conformar un mejor conocimiento del dominio favoreciendo la usabilidad e interoperabilidad entre los sistemas de análisis y la visualización de datos, dando como resultado beneficios para los desarrolladores de servicios y los encargados de la toma de decisiones. Para ello, se emplean técnicas de análisis de *Big Data* que facilitan la extracción del conocimiento para conformar la base de conocimiento ontológico. Posteriormente, se puede sacar provecho del contenido ahí almacenado para la generación de nuevo conocimiento, su visualización y su recuperación.

El término '*Social Big Data*' [4] ha sido acuñado para referirse al análisis de datos masivos provenientes de los medios sociales y abarca áreas como la minería de datos, el aprendizaje computacional, estadística, ontologías y Web semántica, procesamiento del lenguaje natural, etc. En este contexto, las ontologías pueden ayudar en la 'fusión' de datos provenientes de diferentes fuentes para lidiar con la heterogeneidad semántica. En particular, se ha hecho uso de LOD basado en modelos de datos RDF como modelo de datos unificado para combinar, agregar, y transformar datos de fuentes heterogéneas para construir mejores servicios y aplicaciones hibridas para los usuarios. En el análisis de textos en el ámbito del *Social Big Data* también se ha planteado el uso de indexación semántica latente, un método que permite representar los conceptos semánticos de los documentos empleando un conjunto reducido de términos partiendo del principio de que las palabras que se utilizan en el mismo contexto tienden a tener significados similares. Así mismo, ontologías se han empleado para llevar a cabo procesos de análisis de sentimiento más precisos en entornos como Twitter [32].

En [27] se describe un método para la integración de ontologías heterogéneas empleando algoritmos de optimización multiobjetivo partiendo de la motivación de que esta integración permitirá el procesamiento y análisis de datos masivos procedentes de fuentes de información heterogéneas. En resumen, las ontologías facilitan la gestión de la información a nivel de conocimiento y el acceso integrado a fuentes heterogéneas pudiendo emplear diversas técnicas de inferencia y razonamiento para analizar los datos. Sin embargo, antes de que se utilice como nueva infraestructura, hay aspectos que necesitan mejoras [29]: (i) se deben desarrollar algoritmos analíticos más precisos y efectivos que atiendan las necesidades cambiantes de las distintas aplicaciones, y (ii) es preciso encontrar un mecanismo que permita determinar la técnicas de análisis de *Big Data* a emplear sobre las bases de conocimiento en función de las necesidades analíticas de los usuarios finales.

4) Visualización

La exploración y visualización de *Big Data* se ha convertido en un importante desafío de investigación, cuya escalabilidad, funcionalidad y tiempo de respuesta son un requisito vital [28]. Entre los principales desafíos a los que se deben enfrentar los sistemas de visualización modernos se puede destacar (i) el gran tamaño y la naturaleza dinámica de los datos, (ii) la variedad de tareas y usuarios, y (iii) los problemas relacionados con el rendimiento. En el contexto del *Web of Data*, se han desarrollado numerosos sistemas de exploración y visualización que se pueden clasificar en las siguientes categorías [28]: (i) navegadores y sistemas de exploración (por ejemplo, Tabulator), (ii) sistemas de visualización genéricos (por ejemplo, Rhizomer[14]), (iii) sistemas de visualización específicos del dominio, vocabulario y dispositivo (por ejemplo, Map4rdf[15]), (iv) sistemas de visualización basados en grafos (por ejemplo, RelFinder[16]) (v) sistemas de visualización de ontologías (por ejemplo, OWLViz[17]), y (vi) librerías de visualización (por ejemplo, Sgvizler).

Ninguno de los sistemas desarrollados hasta la fecha alcanza los niveles necesarios de rendimiento y escalabilidad (están pensados para manejar pequeños conjuntos de datos). Una de las propuestas más interesantes para la visualización y exploración de datos semánticos es el sistema LDVM (*Linked Data Visualization Model*) [38], que forma parte de la categoría de sistemas de visualización genéricos. LDVM es una adaptación del marco de referencia conceptual DSRM (*Data State Reference Model*) para visualizar RDF y *linked data*. Además, extiende DSRM añadiendo tres conceptos (componentes software reutilizables), a saber, analizadores (que generan la representación en RDF de las fuentes de datos), transformadores (preparar una estructura RDF adecuada en función de la técnica de visualización a emplear), y visualizadores (crean la visualización para el usuario final). El origen de los datos pueden ser fuentes semiestructuradas o no estructuradas.

En resumen, en la actualidad convergen un gran número de aplicaciones de exploración y visualización de datos masivos semánticos. La mayoría de las propuestas se enfocan en tareas o actividades específicas, son pocas las que integran componentes genéricos o de alto grado de abstracción, ya que además trabajan con pequeños grupos de datos y no para todo su volumen, sin duda, un reto que queda pendiente por resolver.

IV. Discusión y Conclusiones

Las soluciones actuales para el análisis de datos (masivos) se enfrentan a numerosas dificultades asociadas, principalmente, con la recolección y almacenamiento integrado de los datos, la búsqueda sobre los mismos, su compartición, análisis y visualización. Las tecnologías de la Web semántica dotan a los datos publicados en la Web de un sustento formal que permite su tratamiento automatizado mediante sofisticadas técnicas de inferencia y razonamiento. Sobre esta base, investigadores en todo el mundo están explorando el modo de aprovechar las sinergias entre el análisis de datos (masivos) y las tecnologías semánticas. En la revisión sistemática de la literatura realizada en este trabajo, se han analizado los beneficios y retos asociados con esta integración. Tras este estudio, es posible dar respuesta a las preguntas de investigación formuladas en la sección II.A.

PI1. ¿Es posible realizar análisis semántico de datos (masivos)?

Se ha comprobado la existencia de multitud de trabajos que exploran las posibilidades de las tecnologías semánticas en el contexto del análisis de datos. Su viabilidad se confirma incluso con el desarrollo de prototipos, herramientas y aplicaciones que integran a estas tecnologías como apoyo y solución a determinados retos que presentan los procesos del análisis de datos. En cuanto al análisis de datos masivos, también hay investigaciones enfocadas en lidiar con los problemas del Big Data con la aplicación de tecnologías semánticas. Sin embargo, no es una tarea sencilla, se deben crear nuevos componentes, e incluso un cambio de paradigma en la forma que se adquiere el conocimiento de los grandes datos.

PI2. ¿En qué etapas del análisis de datos (masivos) tiene sentido aplicar tecnologías semánticas?

Se ha podido comprobar que las tecnologías semánticas son de utilidad en todas las fases del proceso de análisis de datos masivos, desde la recolección y organización de datos, hasta su procesamiento y análisis, y su visualización.

PI3. ¿Cuáles son las principales ventajas de aplicar tecnologías semánticas en el análisis de datos (masivos)?

Los principales beneficios del empleo de semántica en el análisis de datos se pueden agrupar en las siguientes categorías: (i) facilitar la integración de datos de fuentes heterogéneas, (ii) establecer enlaces explícitos entre los datos, (iii) permitir los procesos de razonamiento e inferencia lógica sobre los datos, y (iv) simplificar la exploración de datos vinculados y su visualización.

PI4. ¿Qué retos plantea el uso de tecnologías semánticas en el análisis de datos (masivos)?

Para sacar el máximo provecho de la integración de la semántica con el análisis de datos es necesario superar desafíos como (i) mejorar la eficiencia en la extracción de los datos de diferentes fuentes, (ii) detectar inconsistencias durante el preprocesado de datos para su evaluación y control de calidad, (iii) corregir problemas en el almacenamiento de las tripletas para su recuperación, y (iv) enriquecer la usabilidad de las herramientas de exploración y visualización de datos con el fin de favorecer su uso al usuario común. En relación con el análisis de datos masivos habrá que considerar los siguientes aspecto: (i) ofrecer mayor eficiencia en operaciones de gestión y toma de decisiones en tiempo real, así como incrementar la seguridad en las transacciones en momentos de alta disponibilidad o aumento en la demanda de datos, (ii) desarrollar algoritmos de análisis más precisos y efectivos con base a las necesidades versátiles del usuario final, y (iii) mejorar la propuesta de las aplicaciones de exploración y

[14] http://rhizomik.net/html/rhizomer/
[15] http://oeg-upm.github.io/map4rdf/
[16] http://www.visualdataweb.org/relfinder.php
[17] https://protegewiki.stanford.edu/wiki/OWLViz

visualización con la integración de componentes genéricos y de alto grado de abstracción.

PI5. ¿Cuál es el nivel de madurez de las soluciones que emplean tecnología semántica en el análisis de datos (masivos)?

De acuerdo con esta revisión sistemática de la literatura, los principales aspectos de la integración de semántica y técnicas de análisis de datos (masivos) han sido corroboradas a través de distintas aplicaciones y herramientas validadas en diversos dominios.

A través de este estudio ha quedado patente la sinergia entre las tecnologías semánticas y las soluciones de análisis de datos y Big Data, destacando los beneficios asociados con la recuperación de los datos desde fuentes estructuradas, semi-estructuradas y no estructuradas, su organización como *linked data*, su procesamiento aprovechando el sustento formal de los esquemas ontológicos, y su exploración/visualización aprovechando la estructura enlazada. Quedan, sin embargo, numerosos retos vigentes ligados, fundamentalmente, a la falta de un marco de referencia estándar que establezca una guía sobre el modo en que la semántica puede aplicarse en cada etapa del proceso de análisis de datos.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por la Agencia Española de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER) a través del proyecto KBS4FIA (TIN2016-76323-R).

REFERENCIAS

[1] P. Bihani and S. Patil, "A Comparative Study of Data Analysis Techniques," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 3, no. 2, pp. 95–101, 2014.

[2] D. Airinei and D. Berta, "Semantic Business Intelligence - a New Generation of Business Intelligence," *Inform. Econ.*, vol. 16, no. 2, pp. 72–80, 2012.

[3] A. R. Kim, H.-A. Park, and T.-M. Song, "Development and Evaluation of an Obesity Ontology for Social Big Data Analysis," *Healthc. Inform. Res.*, vol. 23, no. 3, p. 159, 2017.

[4] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Inf. Fusion*, vol. 28, no. 0, pp. 45–59, Mar. 2016.

[5] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*, Second. Elsevier, 2005.

[6] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Commun. ACM*, vol. 54, no. 8, p. 88, 2011.

[7] A. Gómez Vieites and C. Suárez Rey, *Sistemas de información : herramientas prácticas para la gestión empresarial*. RA-MA, 2011.

[8] J. M. Molina López and J. García Herrero, "Técnicas de análisis de datos: Aplicaciones prácticas utilizando Microsoft Excel y Weka," Madrid, 2006.

[9] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263–286, Jan. 2017.

[10] D. Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," *Appl. Deliv. Strateg.*, 2001.

[11] C. Maté Jiménez, "Big data. Un nuevo paradigma de análisis de datos," *Rev. An. Mecánica y Electr.*, vol. 41, no. 6, pp. 10–16, 2014.

[12] M. Beyer and D. Laney, "The Importance of 'Big Data': A Definition," *Gart. Publ.*, no. June, pp. 1–7, 2012.

[13] M. Maier, "Towards a Big Data Reference Architecture," Eindhoven University of Technology, 2013.

[14] L. Joyanes Aguilar, *Big Data, Análisis de grandes volúmenes de datos en organizaciones*, 1st ed. México: Alfaomega Grupo Editor, 2016.

[15] V. Mayer-Schönberger and K. Cukier, *Big data : a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[16] S. A. Kale and S. S. Dandge, "Understanding the Big Data Problems and Their Solutions Using Hadoop and Map-Reduce," *Int. J. Appl. or Innov. Eng. Manag.*, vol. 3, no. 3, pp. 439–445, 2014.

[17] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web.," *Sci. Am.*, vol. 284, no. 5, pp. 34–43, 2001.

[18] W3C, "RDF - Semantic Web Standards," 2014. [Online]. Available: https://www.w3.org/RDF/. [Accessed: 30-Dec-2017].

[19] D. L. McGuinness and F. van Harmelen, "OWL Web Ontology Language - Overview," *W3C*, 2004. [Online]. Available: https://www.w3.org/TR/2004/REC-owl-features-20040210/. [Accessed: 24-Apr-2018].

[20] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data Knowl. Eng.*, vol. 25, no. 1–2, pp. 161–197, Mar. 1998.

[21] M. Barceló Valenzuela, G. G. A. Sánchez Schmitz, and A. Perez-Soltero, "La web semántica como apoyo a la gestión del conocimiento y al modelo organizacional," *Rev. Ing. Informática*, no. 12, p. 4, 2006.

[22] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *Int. J. Semant. Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[23] B. Hu, N. Carvalho, L. Laera, and T. Matsutsuka, "Towards big linked data," in *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services - IIWAS '12*, 2012, vol. 9, no. 4, p. 167.

[24] D. Budgen and P. Brereton, "Performing systematic literature reviews in software engineering," *Proceeding 28th Int. Conf. Softw. Eng. - ICSE '06*, vol. 13, no. 1, p. 1051, Mar. 2006.

[25] A. Siddaway, "What is a systematic literature review and how do I do one?," *Univ. Stirling*, no. Ii, pp. 1–13, 2014.

[26] A. J. Rodríguez-Morales, "La importancia del H index como indicador de la producción y la calidad científica," 2015. [Online]. Available: http://www.redalyc.org/html/849/84943818001/index.html. [Accessed: 30-Dec-2017].

[27] V. Kureychik and A. Semenova, "Combined Method for Integration of Heterogeneous Ontology Models for Big Data Processing and Analysis," vol. 573, Springer, 2017, pp. 302–311.

[28] N. Bikakis and T. Sellis, "Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art," *CEUR Workshop Proc.*, vol. 1558, Jan. 2016.

[29] Q. Bao, J. Wang, and J. Cheng, "Research on Ontology Modeling of Steel Manufacturing Process Based on Big Data Analysis," *MATEC Web Conf.*, vol. 45, p. 4005, Mar. 2016.

[30] T. Shah, F. Rabhi, and P. Ray, "Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions," *Cluster Comput.*, vol. 18, no. 1, pp. 351–367, Mar. 2015.

[31] C. Hu, Z. Xu, Y. Liu, L. Mei, L. Chen, and X. Luo, "Semantic Link Network-Based Model for Organizing Multimedia Big Data," *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 376–387, Sep. 2014.

[32] P. Gupta, "Analysis on Twitter data of automobile domain using ontology," *Int. Educ. Res. J.*, vol. 3, no. 5, pp. 494–496, 2017.

[33] A. Konys, "A Framework for Analysis of Ontology-Based Data Access," vol. 10449, N. T. Nguyen, G. A. Papadopoulos, P. Jędrzejowicz, B. Trawiński, and G. Vossen, Eds. Cham: Springer International Publishing, 2016, pp. 397–408.

[34] D. Lembo, J. Mora, R. Rosati, D. F. Savo, and E. Thorstensen, "Mapping Analysis in Ontology-Based Data Access: Algorithms and Complexity," vol. 9366, 2015, pp. 217–234.

[35] T. Neuböck, B. Neumayr, M. Schrefl, and C. Schütz, "Ontology-Driven Business Intelligence for Comparative Data Analysis," in *Lecture Notes in Business Information Processing*, vol. 172 LNBIP, 2014, pp. 77–120.

[36] V. Sabol, G. Tschinkel, E. Veas, P. Hoefler, B. Mutlu, and M. Granitzer, "Discovery and Visual Analysis of Linked Data for Humans," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8796, 2014, pp. 309–324.

[37] J. Klímek, J. Helmich, and M. Nečaský, "Payola: Collaborative Linked Data Analysis and Visualization Framework," in *ESWC (Satellite Events)*, vol. 7955, no. 257943, 2013, pp. 147–151.

[38] J. M. Brunetti, S. Auer, R. García, J. Klímek, and M. Nečaský, "Formal Linked Data Visualization Model," in *Proceedings of International Conference on Information Integration and Web-based Applications &*

Services - IIWAS '13, 2013, vol. 2, no. 257943, pp. 309–318.
[39] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, Jul. 2012.
[40] A. Musetti *et al.*, "Aemoo: Exploratory search based on knowledge patterns over the semantic web," *Semant. Web Chall.*, 2012.
[41] V. Presutti, L. Aroyo, A. Adamou, B. Schopman, A. Gangemi, and G. Schreiber, "Extracting core knowledge from Linked Data," *Proc. Second Int. Conf. Consum. Linked Data*, vol. 782, no. COLD'11, pp. 37–48, 2011.
[42] I. Kupershmidt *et al.*, "Ontology-Based Meta-Analysis of Global Collections of High-Throughput Public Data," *PLoS One*, vol. 5, no. 9, p. e13066, Sep. 2010.
[43] F. Prosdocimi, B. Chisham, E. Pontelli, J. D. Thompson, and A. Stoltzfus, "Initial implementation of a comparative data analysis ontology," *Evol. Bioinforma.*, vol. 2009, no. 5, pp. 47–66, 2009.
[44] A. G. Nuzzolese, V. Presutti, A. Gangemi, S. Peroni, and P. Ciancarini, "Aemoo: Linked Data exploration based on Knowledge Patterns," *Semant. Web*, vol. 8, no. 1, pp. 87–112, 2017.
[45] G. Ganapathy and S. Sagayaraj, "Automatic Ontology Creation by Extracting Metadata from the Source code," *Glob. J. Comput. Sci. Technol. GJCST Classif. D*, vol. 10, no. 14, pp. 12–18, 2010.
[46] J. M. Gimenez, J. D. Fernandez, and M. A. Martinez, "A MapReduce-based Approach to Scale Big Semantic Data Compression with HDT," *IEEE Lat. Am. Trans.*, vol. 15, no. 7, pp. 1270–1277, 2017.
[47] J. O. Chan, "Big Data Customer Knowledge Management," *Commun. IIMA*, vol. 14, no. 3, pp. 45–56, 2014.

Héctor Hiram Guedea Noriega received the B.Eng. and M.Tech. degrees from University of Colima, México. He is currently PhD student in Computer Science at the University of Murcia, Spain. He has more than 10 years of experience in Web development with advanced knowledge in front-end and back-end technologies. He is working as Senior Software Engineer and Digital Management for Secretariats of State of Colima and a software development company from Ontario, Canada. His main research interests are Semantic Web technologies, Digital Marketing, Social Semantic Web and Big Data.

Francisco García Sánchez Francisco Garcia-Sanchez received the B.E., M.Sc., and Ph.D. degrees in computer science from the University of Murcia, Murcia, Spain. He is currently an Associate Professor in the Department of Informatics and Systems, University of Murcia. His main research interests are Semantic Web-based applications, Natural Language Processing, Semantic Service-Oriented Architectures, Social Semantic Web, and the application of agent technologies. He has conducted a number of research stays in world-leading research institutes in Ireland, Austria, the United States and Australia, and has published over 60 articles in journals, conferences, and book chapters. He has taken part in various research projects related to the application of Semantic Web technologies to real world challenges as both principal investigator and researcher.