INSTITUTO POLITÉCNICO NACIONAL

Centro de Investigación en Computación

Integración semántica de fuentes heterogéneas de datos meteorológicos con base en datos vinculados.

Tesis

q u e p a r a o b t e n e r e l g r a d o d e

Doctorado en Ciencias de la Computación

p r e s e n t a

Luis Cabrera Rivera

Asesores:

Dr. Miguel Jesús Torres Ruiz

Dr. Luis Manuel Vilches Blázquez

México, D.F., Noviembre del 2018

Resumen

En la actualidad se generan una gran cantidad de información meteorológica y de calidad del aire por parte de las instituciones gubernamentales y de fuentes voluntarias, sin embargo, al provenir la información de diferentes fuentes, se genera un problema de heterogeneidad en la información, provocando que, a pesar de tener una gran cantidad de datos, estos no se exploten de manera conjunta.

Por todo lo anterior, en este trabajo se propone una metodología capaz de integrar la información semánticamente, que proviene de las fuentes de datos oficial y voluntarias, ya sean de tipo estático o dinámico, utilizando los principios de *Linked Data* para enriquecer la información.

Esta metodología esta conformada por cuatro bloques: 1) *Recolección de datos*, 2) *Pre procesamiento*, 3) *Integración Semántica* y 4) *Análisis*. El primer componente ese centra en la recolección de datos, que se lleva a cabo de dos maneras diferentes: estática y dinámica; El segundo componente es el asociado con el pre-procesamiento de los datos que han sido recuperados por el componente anterior; El tercer componente es el módulo de integración semántica, este componente recoge el detalle de la implementación de la red ontológica, además de la población de la red ontológica a través de un proceso automático de generación de RDF y el establecimiento de conexiones a la nube de *Linked Data*. Finalmente, el componente de análisis que permite explotar los datos integrados semánticamente a través de diversas operaciones de análisis.

Como resultado se obtiene un repositorio de datos integrados semánticamente, el cual es explotable para su análisis. Como caso de estudio se tomó Ciudad de México, por las diversas estaciones tanto oficiales como voluntarias que existen en ella.

Abstract

Currently a lot of meteorological and air quality information are generated by government institutions and voluntary sources, however, all the information comes from different sources, generates a heterogeneity problem on information, this causing that despite a large amount of data, this without using both sources at same time.

Because of the previously, in this work propose a methodology capable of integrate semantically the information, from the official and voluntary sources, whether static or dynamic type, this using the Linked Data principles to enrich the retrieved information.

This methodology is composed by four components: 1) Data recollection, 2) Pre-processing, 3) Semantic integration and 4) Analysis. The first component focuses in recollecting data, from two different ways: static and dynamic; The second component focuses is associate with pre-processing from data retrieved from last component. The third component focuses in the semantic integration, this recollect all details from the develop of the ontological network, also of the population of the ontology network through an automatic process of generation of RDF files and make all connections to *Linked Data Cloud*. Finally, the fourth block focuses in exploit the integrate semantic data through different analysis operations.

As result is obtained a semantic integrated data repository, that can be exploit for analysis. As case of study is considering the Mexico City, because all different stations both official and voluntary that exist in the city.

Índice

Índice de Figuras

Índice de Tablas

CAPÍTULO 1 INTRODUCCIÓN

Durante años las personas han recabado y utilizado datos sobre el clima, con la finalidad de llevar a cabo análisis que permitan generar predicciones sobre el comportamiento del clima en el futuro. Al mismo tiempo estos datos se han utilizado para reconocer ciertos patrones que permiten crear una clasificación de fenómenos meteorológicos conforme a variables como fuerza del viento, humedad, presión atmosférica y temperatura, entre otros.

Actualmente con el problema del cambio climático las Naciones Unidas(UN) han lanzado 17 objetivos para transformar nuestro mundo, siendo el objetivo 13[1] trata sobre las repercusiones que tienen el cambio climático sobre nuestras vidas diarias, según la UN de seguir el rumbo actual en estas siglo el aumento en la temperatura podía ser de 3 grados, lo que repercutiría en la sociedad en general sin importar país alguno, a continuación se muestran datos específicos se la situación hasta el 2017 que la UN proporciona los siguientes:

- **Entre 1880 y 2012, la temperatura media mundial aumentó 0,85 grados centígrados.** Esto quiere decir que por cada grado que aumenta la temperatura, la producción de cereales se reduce un 5% aproximadamente. Se ha producido una reducción significativa en la producción de maíz, trigo y otros cultivos importantes, de 40 megatones anuales a nivel mundial entre 1981 y 2002 debido a un clima más cálido.

- **Los océanos se han calentado, la cantidad de nieve y de hielo ha disminuido, y ha subido el nivel del mar.** Entre 1901 y 2010, el nivel medio del mar aumentó 19 cm, pues los océanos se expandieron debido al calentamiento y al deshielo. La extensión del hielo marino del Ártico se ha

[1] https://www.un.org/sustainabledevelopment/es/climate-change-2/

reducido en los últimos decenios desde 1979, con una pérdida de hielo de 1,07 millones de km2 cada decenio

- **Dada la actual concentración y las continuas emisiones de gases de efecto invernadero, es probable que a finales de siglo el incremento de la temperatura mundial supere los 1,5 grados centígrados en comparación con el período comprendido entre 1850 y 1900 en todos los escenarios menos en uno.** Los océanos del mundo seguirán calentándose y continuará el deshielo. Se prevé una elevación media del nivel del mar de entre 24 y 30 cm para 2065 y entre 40 y 63 cm para 2100. La mayor parte de las cuestiones relacionadas con el cambio climático persistirán durante muchos siglos, a pesar de que se frenen las emisiones

- Las emisiones mundiales de dióxido de carbono (CO2) han aumentado casi un 50% desde 1990

- Entre 2000 y 2010 se produjo un incremento de las emisiones mayor que en las tres décadas anteriores

Por tal motivo los gobiernos de cada país se han ocupado en monitorear los efectos en el clima que han surgido a causa del calentamiento global. Esto implica que se requiere una gran cantidad de datos de alta calidad, por lo cual la calibración de las estaciones base es de alta importancia, así como el área de cobertura de cada una de estas.

Sin embargo, el acceso a este tipo de datos para su análisis suponía una gran inversión monetaria, ya que solo las instituciones gubernamentales eran las únicas que podían consultar la información bajo permiso, para la sociedad en general el acceso a esta información requiere pagar precios exorbitantes.

De tal modo que varias personas ha adquirir equipo especializado con los cuales recolectan la información y la ponen a disposición de la gente que las necesita (Goodchild, Li, 2012), mediante el uso de dispositivos como las estaciones

bases meteorológicas generan una gran cantidad de nuevos datos y las comparten por medio de Internet, usando los diferentes sitios Web disponibles. Esta práctica es definida como Información Geográfica Voluntaria (VGI) y (Goodchild, 2007) la define como el aprovechamiento de las herramientas para crear, reunir y difundir información geográfica proporcionada voluntariamente por los individuos.

Dentro del marco de los fenómenos meteorológicos también se ha empezado a recopilar información meteorológica por parte de los usuarios o pequeñas asociaciones no gubernamentales, como son los sitios web AccuWeather[2], WeatherUnderground[3], GEONAMES[4], Foreca[5], Weatherbug[6], 3TIER[7], Data.gov.sg[8], Aeris Weather[9] entre otros. Estos sitios se dedican a recopilar información de estaciones bases voluntarias a tres del mundo y proporcionar predicciones meteorológicas, algunas de ellas como WeatherUnderground o AccuWeather visualizan la información en un mapa, por otro lado, sitios como Aeris Weather, Foreca GEONAMES recopila además información sobre calidad del aire.

Según (Goodchild, Li, 2012) la información que se genera a partir del VGI ha probado ser de gran utilidad debido a la rapidez con las que se genera información geográfica detallada a un bajo costo cambiando el paradigma tradicional donde el gobierno únicamente generaba y proporcionaba la información. Además acorde a los trabajos de (Sui, Elwood, Goodchild, 2012) y (Fischer, 2012) el VGI cumple las 5 Vs que plantea Big Data según el trabajo de (Demchenko, De Laat, Membrey, 2014), esto debido a que se generan grandes volúmenes de información (*Volumen*) a una gran velocidad debido al continuo censado de las diferentes estaciones base de donde se toman mediciones (*Velocity*), dando como resultado los diferentes

[2] https://www.accuweather.com
[3] https://www.wunderground.com/
[4] http://www.geonames.org/export/
[5] http://www.foreca.es
[6] https://www.weatherbug.com/
[7] https://www.3tier.com/
[8] https://data.gov.sg/
[9] https://www.aerisweather.com/

formatos de archivos dependiendo de las fuentes (*Variety*). Además, cada una de las medidas presenta diferentes grados de precisión con respecto a las mediciones oficiales (*Veracity*), dándole un valor agregado como información complementaria con respecto a los datos obtenidos de las instituciones gubernamentales o, incluso, permitiendo generar información donde las instituciones gubernamentales no tienen estaciones meteorológicas (*Value*). Sin embargo, la información voluntaria proporcionada al ser generada por los usuarios que no son expertos en el área no siempre son precisas en comparación con las fuentes oficiales, debido a varios factores, uno es el acceso a equipos con sensores de alta precisión, teniendo que usar equipos de bajo costo con sensores menos precisos, otro factor puede ser la mala calibración a los equipos realizada por los usuarios, debido a eso la información proporcionada por los usuarios puede carecer de calidad y precisión, problemas inherentes al VGI (Sui, Elwood, Goodchild, 2012).

Otro problema común es que la información meteorológica se genera en grandes volúmenes y en tiempo real por lo que es complicado analizarla y conocer la calidad y coherencia de esta información, por tal motivo es importante realizar una comparación entre los datos meteorológicos proporcionados por las instituciones gubernamentales, como Comisión Nacional del Agua(CONAGUA) y la Secretaria de Marina(SEMAR) a nivel federal y a nivel local como la Red Automática de Monitoreo Atmosférico(RAMA) y la Red Meteorológica y Radiación Solar (REDMET) de la CDMX, así como con los datos recopilados por los voluntarios mediante sus estaciones meteorológicas, con la finalidad de realizar una integración de ambas fuentes de información y poder generar un repositorio de datos a los que pueda tener acceso el público en general.

1.1 Descripción del problema

En la actualidad se generan una gran cantidad de información meteorológica por parte de las instituciones gubernamentales y de fuentes voluntarias, esto debido a la velocidad en la que se genera la información, en el caso de las fuentes oficiales

como RAMA y REDMET proporcionan lecturas cada hora del día de cada una de las estaciones base que monitorean; En el caso de las fuentes voluntarias cada sitio monitorea las estaciones meteorológicas y calidad del aire proporciona en determinados lapsos de tiempos actualizaciones de las medidas. Al recuperar información oficial y voluntaria, se genera una heterogeneidad de la información de dos clases, la primera siendo el formato de salida de la respuesta y el segundo el diferente etiquetado de las variables que monitorean.

Debido a esta heterogeneidad, al momento de extraer la información, tanto de los archivos como respuestas recuperadas de ambos tipos de fuentes de datos existen discrepancias entre las etiquetas asignadas a las variables que las estaciones base miden. En cuanto al formato de salida, las fuentes de datos voluntarias proporcionan respuestas a sus servicios en diferentes formatos, por ejemplo: archivos JSON y/o XML. Estos formatos difieren a los archivos que proporcionan las fuentes oficiales, que en su mayoría son archivos CSV.

Otra parte del problema se centra en la disponibilidad de la información actualizada de las fuentes oficiales, actualmente gracias a la iniciativa de datos abiertos del gobierno federal, las instituciones gubernamentales proporcionan información de manera libre para el uso de la gente, sin embargo, esta información no siempre esta actualizada ya que depende de cada institución actualizar cada uno de sus repositorios.

Por último un problema inherente al la información voluntaria es la falta de conocimiento sobre la calidad de la información como menciona (Sui, Elwood, Goodchild, 2012), esto por múltiples factores como la precisión de los equipos que monitorean, la falta de conocimiento en la calibración de los equipos entre otras, a diferencia de las fuentes oficiales donde gracias a la inversión del gobierno se pueden adquirir equipos de última generación con sensores de mayor precisión y que son monitoreados por personas con conocimiento en el manejo de los mismos. Lo que provoca que se le de un mayor grado de confiabilidad a los datos producidos por las fuentes oficiales.

1.2 Solución propuesta

Con base en el planteamiento del problema se propone diseñar e implementar una metodología que permita recuperar e integrar semánticamente múltiples y heterogéneas fuentes de información oficiales y voluntarias que presentan características estáticas (archivos CSV) y dinámicas (servicios web, APIs, etc.) por medio de una integración basada en ontologías como se ve en el trabajo de (Wache, et al, 2001), con la finalidad de generar un gran repositorio de datos semánticos publico, que sirva como base para realizar análisis sobre ellos.

1.3 Objetivos Generales

El objetivo general del trabajo consiste en proporcionar una metodología la cual permita integrar semánticamente fuentes heterogéneas de datos meteorológicos con base en datos vinculados.

1.4 Objetivos específicos

- Desarrollar un servicio de recuperación de información de los diferentes tipos de fuentes disponibles ya sean dinámicas o estáticas.

- Diseñar un módulo de pre-procesamiento capaz de monitorear de *streamings* de datos.

- Diseñar y construir una red ontológica que permita modelar el conocimiento relacionado con variables meteorológicas y de calidad del aire.

- Diseñar e implementar un módulo de integración semántica que permita la transformación de los datos originales a archivos RDF.

- Diseñar e implementar un módulo encargado de la comparación de calidad de información y de predicción de calidad del aire.

1.5 Justificación del trabajo

Este trabajo brinda una opción enfocada a la integración de las fuentes heterogéneas de información meteorológica y de calidad del aire, ya sean voluntarias u oficiales, empleando un enfoqué semántico al realizar una integración basada en una red de ontologías para estandarizar las mediciones recuperadas, con la finalidad de enriquecer la información de fuentes oficiales con las voluntarias, obteniendo un repositorio de datos actualizado y publico.

1.6 Alcances y limitaciones

La metodología será capaz de recuperar de las diversas fuentes información meteorológicas y de calidad el aire, para ser extraída y pre procesada la información, posteriormente enviada por un *streaming* de datos para su integración mediante la red de ontologías diseñada y al mismo tiempo auto poblar la red, por último, esta información pasará a ser usada como repositorio de datos del análisis sobre predicciones y calidad de datos. Como limitantes de este trabajo se encuentran:

- La falta de *streaming* constantes de las diferentes fuentes de información.

- La falta de información disponible y actualizada por parte de las fuentes de datos oficiales.

- La dependencia a la disponibilidad de los servicios web y APIs, así como as sus intervalos de actualización.

- La metodología está limitada a las estaciones bases proporcionadas por los servicios web, así como por las ubicaciones en las que se encuentran las estaciones base de las dependencias gubernamentales.

- La metodología está limitada en cuanto a la disponibilidad de la localización geográfica de las estaciones bases para realizar operaciones espaciales sobre estas.

1.7 Aportaciones

1.7.1 Aportaciones científicas

Como aportación científica se obtiene una metodología para la integración semántica de información meteorológica y de calidad del aire, además de una red ontológica diseñada para la integración de este tipo de información.

1.7.2 Aportaciones tecnológicas

Como aportación tecnológica se obtiene una *framework* capaz de recuperar, procesar, integrar y analizar la información proveniente de las diferentes fuentes de datos voluntarias y oficiales.

1.7.3 Aportaciones académicas

Como aportación académica se obtiene repositorio semántico de información meteorológica y de calidad del aire actualizado y de libre acceso.

1.8 Organización de la Tesis

Este trabajo está compuesto por los siguientes capítulos:

- Capítulo 2. Este capítulo está centrado sobre el marco teórico de la tesis, donde se mencionan los conceptos relevantes asociados a este trabajo, tales como: metodologías para la creación de ontologías, herramientas usadas, entre otros.

- Capítulo 3. Este capítulo presenta el estado del arte, compuesto por trabajos relacionados en un sentido científico y tecnológico, sobre Big Data, ingeniería ontológica y VGI.

- Capítulo 4. Este capítulo describe la metodología propuesta con base en los conceptos descritos en el marco teórico, así como definiciones de la arquitectura.

- Capítulo 5. Este capítulo presenta los experimentos realizados y los resultados obtenidos, los cuales servirán para probar la metodología propuesta en este trabajo.

- Capítulo 6. Este capítulo recoge las principales contribuciones y conclusiones obtenidas en el proceso de investigación desarrollado en el trabajo, así como las propuestas para realizar trabajo futuro.

CAPÍTULO 2 MARCO TEÓRICO

En este capítulo, se describen algunos de los términos y herramientas que son utilizados durante el desarrollo de la metodología propuesta. Primeramente, se presenta una descripción general de la definición de ontología y algunos conceptos importantes de la ingeniería ontológica como son las metodologías de creación. Posteriormente, se describen de manera general algunos conceptos importantes relacionados con la definición de *Linked Data* como son, el estándar *RDF* y *SPARQL*. Finalmente, una descripción de *Big Data y* algunas de las herramientas empleadas para su manipulación.

2.1 Ontología

En la literatura actualmente existen varias definiciones de lo que es una ontología, en este trabajo describimos las dos definiciones que consideramos de mayor relevancia para nuestro caso de estudio:

1. "Una ontología define los términos y relaciones básicas que componen el vocabulario de un área temática, así como las reglas para combinar términos y relaciones que definen la extensión del vocabulario" (Neches, 1991).
2. “Una ontología es una especificación explicita de una conceptualización”. (Gruber, 1995).

Para el desarrollo de una ontología, se necesita definir un dominio de discurso y un marco teórico, que sentarán las bases para determinar el vocabulario del cual resultarán los conceptos, las relaciones, las instancias, las constantes, los atributos, axiomas y las reglas. En la tabla 2.1, se define de acuerdo al trabajo de [(Gruber, 1995), (Corcho et al., 2005)] los siguientes conceptos:

Tabla 2.1, Conceptos definidos en el trabajo de [(Gruber, 1995), (Corcho et al., 2005)].

Conceptos	Definición
Conceptos	Son objetos o entidades, considerados desde un punto de vista amplio. Los conceptos de una ontología están normalmente organizados en taxonomías en las cuales se pueden aplicar mecanismos de herencia.
Relaciones	Las relaciones representan un tipo de asociación entre conceptos del dominio. Si la relación une dos conceptos se denomina relación binaria. Una relación binaria relevante es Subclase-de, que se utiliza para construir taxonomías de clase, como se ha especificado anteriormente.
Instancia	Se utilizan para representar individuos en la ontología. Las relaciones también se pueden instanciar.
Constantes	Son valores numéricos que no cambian en un largo período de tiempo.
Atributos	Los atributos describen propiedades, se pueden distinguir dos tipos de atributos de instancia y de clase. • Los atributos de instancia describen propiedades de las instancias de los conceptos, en las cuales toman su valor, éstos se definen en un concepto y se heredan a sus subconceptos e instancias. • Los atributos de clase describen conceptos y toman su valor en el concepto en el cual se definen. Estos atributos no se heredan ni a los subconceptos ni a las instancias.

Axiomas	Los axiomas son expresiones lógicas siempre verdaderas que suelen utilizarse para definir restricciones en la ontología.
Reglas	Las reglas se utilizan normalmente para inferir conocimientos en la ontología, tales como valores de atributos, instancias de relaciones, etc.

De acuerdo al trabajo de (Buriano et al., 2006), para el diseño de una ontología de deben cumplir ciertos criterios, que se describen a continuación:

1. **Claridad y objetividad**.- La ontología debe proporcionar el significado de términos definidos entregando definiciones objetivas y documentadas en lenguaje natural.
2. **Completitud**.- Una definición expresada por una condición necesaria y suficiente es preferida por una definición parcial.
3. **Coherencia**.- Permite que se puedan realizar inferencias y que éstas sean consistentes con las definiciones ya preestablecidas.
4. **Maximiza la extensibilidad monotónica**.- Los términos generales nuevos o especializados deben incluirse en la ontología de tal forma que no requiera revisión de definiciones existentes.
5. **Mínimo compromiso ontológico**.- Hace pocas afirmaciones acerca del mundo a ser modelado, lo que significa que la ontología debe ser lo más específica posible con el significado de sus términos, dando la libertad a la ontología para especializar e instanciar.
6. **Principio de distinción ontológica**.- Las clases de una ontología deben ser disjuntas. El criterio utilizado para aislar las propiedades principales consideradas a ser invariantes para una instancia de una clase se llama criterio de identidad.
7. **Diversificación de jerarquías**.- Si el conocimiento es suficiente es representado en la ontología y con muchas formas o criterios de clasificación,

para que facilite el introducir nuevos conceptos y heredar propiedades de diferentes puntos de vista.

8. **Modularidad**.- Minimiza el acoplamiento entre módulos.
9. **Minimizar la distancia semántica entre conceptos hermanos**.- Los conceptos similares son agrupados y representados como subclases de una clase y deben definirse utilizando las mismas primitivas, mientras que los conceptos menos similares son separados en la jerarquía.
10. **Estandarización de nombres.-** Se lleva a cabo la estandarización, para evitar inconsistencias en la ontología, así como la confusión al momento de realizar inferencias en la misma.

En el trabajo de [(Uschold, Gruninger, 1996), (Roche, 2003)], se describe una categorización con base en el grado de formalidad de las ontologías, obteniendo cuatro categorías que se detallan en los siguientes puntos:

- **Alta informalidad**.- Son aquellas ontologías que están expresadas en lenguaje natural, un ejemplo son los glosarios.
- **Semi-informalidad**.- Estas ontologías estas estructuradas y restringidas por el lenguaje natural y son usadas en orden de reducir la ambigüedad.
- **Semi-formal**.- Estas ontologías están expresadas en un lenguaje artificial definido formalmente como los lenguajes de marco.
- **Riguroso**.- Son ontologías que están precisamente definidas con semántica formal, por ejemplo, los lenguajes basados en lógica.

En el trabajo de (C. Allocca et al., 2009) se define una red ontológica como una colección de ontologías relacionadas entre sí a través de una variedad de relaciones diferentes, como mapeo, modularización y control de versiones, entre otros. Y pueden ser clasificadas, de acuerdo al tipo de conocimiento que será transmitido por la ontología, como se describe en el trabajo de [(Roche, 2003), (Guarino, 1995)]:

1. **Ontologías genéricas**.- Estas ontologías abarcan conceptos generales definidos independientemente del dominio de la aplicación y puede ser usada en varios dominios.

2. **Ontologías de dominio**.- Estas ontologías están definidas para un dominio en particular y contienen conceptos genéricos del mismo, lo que permite que sean reutilizadas en tareas diferentes que están relacionadas con el dominio de discurso.
3. **Ontologías de aplicación**.- Las ontologías de aplicación, usan el conocimiento especifico para una tarea en particular, que incluye conocimiento especifico de expertos para la aplicación, por lo general estas ontologías no se pueden reutilizar.
4. **Meta-ontología**.- Esta ontología especifica la representación del conocimiento usada para definir lo conceptos del dominio y ontologías genéricas

La Ingeniería Ontológica, surgida de la Web Semántica[10], proporciona los requisitos necesarios para mejorar las búsquedas de información. Esta mejora se debe a que en lugar de utilizar palabras clave en los procesos de búsqueda, se centra en los significantes de los conceptos, es decir, en la semántica de la información. De esta manera, se obviará la asunción de que los datos deben ser entendidos exclusivamente por los usuarios y se pasará a un proceso de entendimiento recíproco entre hombre y máquina, en la que las máquinas pasarán a "*comprender*" los datos que procesan, actuando sin la necesaria y continuada supervisión actual, comentan en su trabajo Blázquez et al., 2006).

2.2 Metodologías para la construcción de ontologías

Para la implementación de una ontología es primordial seleccionar la metodología que permita representar el objetivo y uso que se necesita. (Guzmán et al., 2012) describe diversas metodologías existentes en la literatura para diseño e implementación, que se enlistan a continuación:

[10] http://www.w3.org/2001/sw/

1. **CYC**.- Fue publicada por Lenat y Guha en 1990 y describe de manera general los pasos para la construcción de las ontologías. El primer paso, consiste en la extracción manual del conocimiento común que esta implícito en diversas fuentes, para después cuando se tenga suficiente conocimiento en la ontología adquirir nuevo conocimiento común usando herramientas del procesamiento de lenguaje natural o aprendizaje computacional.

2. **Uschold y King**.- Propuesta en el trabajo de Uschold y King en 1995, se emplea el Modelo Enterprise, donde recrean una serie de pasos que plasman y especifican los conocimientos obtenidos sobre un dominio especifico, centrando sus esfuerzos en la forma en la cual representar conocimientos.

3. **Grüninger y Fox**.- Por otro lado, en el trabajo de Grüninger y Fox en 1995, es desarrollada en paralelo de la metodología del Uschold y King y se empleó para construir las ontologías del proyecto *TOVE* (Toronto Virtual Enterprise). Este enfoque utiliza un conjunto de preguntas en lenguaje natural, llamadas cuestiones de competencia, que determinan el ámbito de la ontología y extrae los conceptos principales, sus propiedades, relaciones y axiomas.

4. **Kactus**.- Propuesta en 1996, usa el dominio de las redes eléctricas para desarrollar ontologías como parte del proyecto *Spirit KACTUS*. Esta metodología emplea una base de conocimiento por medio de un proceso de abstracción.

5. **Methontology**.- Desarrollada por (Fernández et al., 1997), esta metodología define actividades para la planificación del proyecto, la calidad del resultado, la documentación, etc. La metodología esta compuesta por once tareas definidas en el trabajo de (Corcho et al., 2005) mostradas en la Figura 2.1.

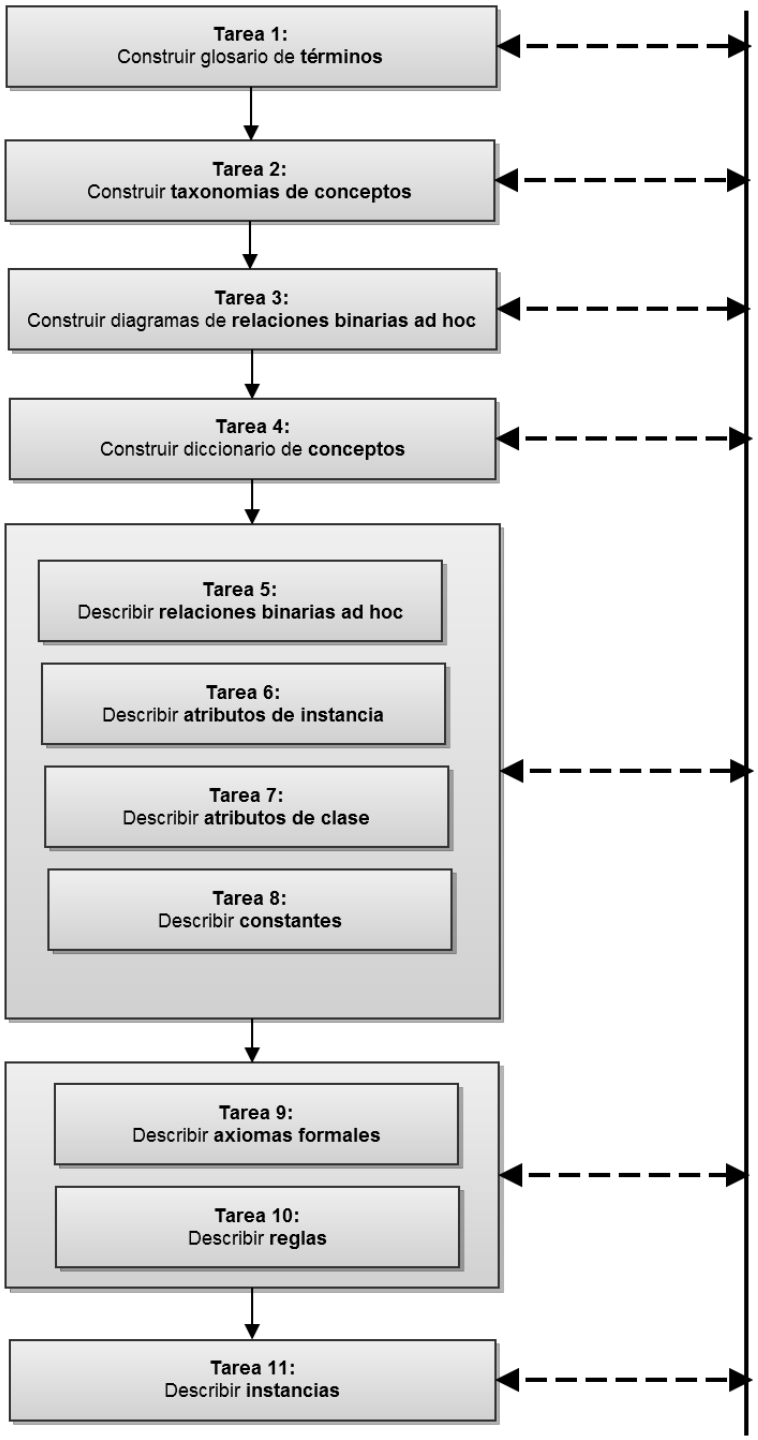


Figura. 2.1. Tareas de la actividad de Conceptualización según METHONTOLOGY (Corcho et al., 2005).

Sensus.- Definida en 1997, como un nuevo método para la construcción de ontologías, la cual constituye un enfoque *top-down* para derivar ontologías específicas del dominio a partir de grandes ontologías. Sensus identifica un conjunto de términos semilla que son relevantes para el dominio particular. Dichos términos se enlazan manualmente a una ontología de amplia cobertura, con lo cual el usuario puede seleccionar los términos más relevantes y así acotar la ontología *Sensus*. Posteriormente, el algoritmo regresa un conjunto de términos estructurados jerárquicamente para describir un dominio, que puede ser usado como base de conocimiento.

ON-TO- KNOWLEDGE.- En su trabajo, (Sure et al., 2003) desarrolla esta metodología que aplica ontologías a la información disponible electrónicamente, con el objetivo de mejorar la calidad de la gestión de conocimiento en organizaciones grandes y distribuidas. Además, incluye la identificación de metas que deberían ser conseguidas por herramientas de gestión de conocimiento, y se basan en el análisis de escenarios de uso y en los diferentes papeles desempeñados por trabajadores de conocimiento y accionistas en las organizaciones.

NeOn.- Propuesta por (Suárez et al., 2012), la metodología NeOn está diseñada para la construcción de redes de ontologías, basada en escenarios que se apoyan en los aspectos de colaboración de desarrollo de ontologías, además de en la reutilización y evolución dinámica de las redes de ontologías en entornos distribuidos.

Las claves de la Metodología *NeOn,* son un conjunto de nueve escenarios para la construcción de ontologías y redes de ontologías, como se muestra en la Figura 2.2, haciendo hincapié en la reutilización de los recursos ontológicos y no ontológicos, la reingeniería y la fusión, y teniendo en cuenta la colaboración y el dinamismo. El Glosario de Procesos y Actividades identifica y define aquellos procesos y actividades involucrados en el desarrollo de las redes de ontologías.

Directrices metodológicas para diferentes procesos y actividades del proceso de desarrollo de la ontología de la red, tales como la reutilización y la reingeniería

de los recursos ontológicos y no ontológicos, la especificación de los requisitos de la ontología, la localización de la ontología, la programación, etc. Todos los procesos y actividades se describen con (a) una tarjeta llena, (b) un flujo de trabajo, y (c) ejemplo.

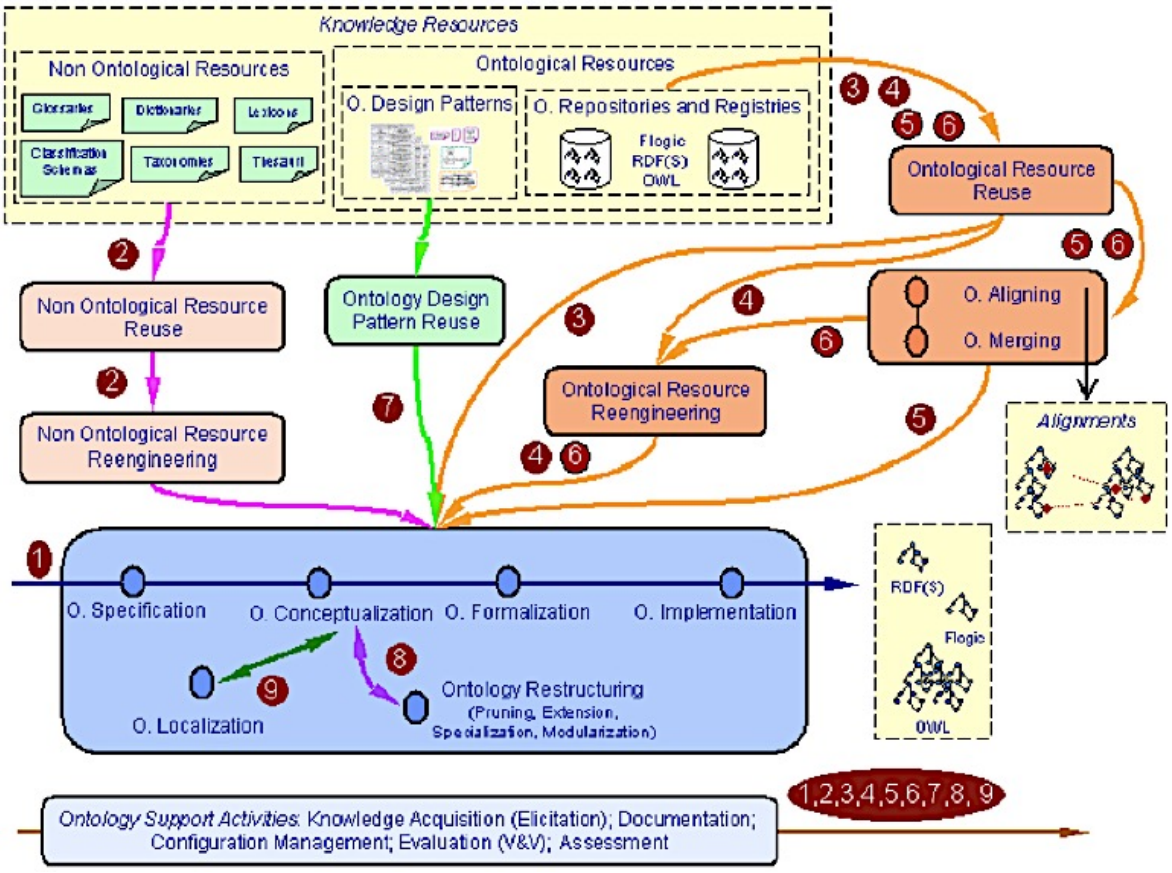


Figura. 2.2. Escenarios para la construcción de ontologías y ontologías de red, usando NeOn (Suárez et al., 2012).

Escenario 1: Desde la especificación de la aplicación. La red de ontologías es desarrollada sin volver a utilizar los recursos existentes. Los desarrolladores especifican los requisitos de la ontología. Posteriormente, se asesoran para llevar a cabo una búsqueda de recursos potenciales para ser reutilizados. A continuación, la actividad de planificación se debe realizar, y los desarrolladores deben seguir el plan.

Escenario 2: La reutilización y reingeniería de los recursos no ontológicos (NOR). Los desarrolladores deben llevar a cabo el proceso de

reutilización NOR para decidir, conforme a los requisitos de la ontología, que *NORs* pueden ser reutilizados para construir la red de la ontología. A continuación, los *NORs* seleccionados deben volver al proceso de re-ingeniería ontológica.

Escenario 3: La reutilización de los recursos ontológicos. Los desarrolladores utilizan recursos ontológicos (ontologías como un conjunto de módulos ontológicos, y/o declaraciones) para construir redes de ontologías.

Escenario 4: La reutilización y re-ingeniería de los recursos ontológicos. Los desarrolladores de ontologías reutilizan los recursos y reorganizar los recursos ontológicos.

Escenario 5: La reutilización y la fusión de los recursos ontológicos. Este escenario se produce cuando varios recursos ontológicos en el mismo dominio se seleccionan para su reutilización, y los desarrolladores desean crear un nuevo recurso ontológico con los recursos seleccionados.

Escenario 6: Reutilización, la fusión y re-ingeniería de los recursos ontológicos. Los desarrolladores de ontologías reutilizan, combinan y reorganizan los recursos-ontológicos. Este escenario es similar al Escenario 5, pero en este caso los desarrolladores deciden reorganizar el conjunto de recursos combinados.

Escenario 7: Reutilización de los patrones de diseño de ontologías (ODPs). Los desarrolladores de ontologías acceden a repositorios de reutilización ODPs.

Escenario 8: Reestructuración de recursos ontológicos. Los desarrolladores de ontologías reestructuran (modularizan, podan, extienden y/o especializan) recursos ontológicos que deben integrarse posteriormente en la red de ontologías.

Escenario 9: Localización de recursos ontológicos. Los desarrolladores de ontologías adaptan una ontología a otros idiomas y la cultura de las comunidades, obteniendo así una ontología multilingüe.

2.3 Linked Data

Linked Data es la forma que tiene la Web Semántica de vincular distintos datos que están distribuidos en la Web, de forma que se referencian de la misma forma que lo hacen los enlaces de las páginas Web.

La Web Semántica no se trata únicamente de las publicaciones de datos en la Web, sino que éstos se pueden vincular a otros, de forma que las personas y las máquinas puedan explorar la web de los datos, permitiendo llegar a información relacionada que se hace referencia desde otros datos iniciales. A diferencia de la web del hipertexto, donde los enlaces son relaciones entre puntos de los documentos escritos en *HTML*, los datos enlazan cosas arbitrarias que se describen en *RDF*.

Linked Data se basa en la aplicación de ciertos principios básicos y necesarios, que documentarán el crecimiento de la Web, tanto a nivel de los documentos *HTML* (vista clásica de la Web), como a nivel de los datos expresados en *RDF* (vista de la Web Semántica)

1. Usar *URIs* para identificar las cosas
2. Usar *URIs* http
3. Ofrecer información sobre los recursos usando *RDF*
4. Incluir enlaces a otros *URIs*

Al nombrar los conceptos o cosas mediante *URIs*, se ofrece una abstracción del lenguaje natural y así se consigue evitar ambigüedades y ofrecer una forma estándar y única para referirnos a cualquier recurso. Ya que existen muchos esquemas de *URIs*, se pretende el uso de *URIs* sobre http para asegurar que cualquier recurso pueda ser buscado y accedido en la *Web*. Debe tenerse en cuenta que los *URIs* no son sólo direcciones, son identificadores de recursos (Bizer, C., Heath, T., & Berners-Lee, T., 2009).

Una vez que se busca y se accede a un recurso identificado mediante una *URI http*, se debe obtener información útil sobre dicho recurso, representada mediante descripciones estándares en *RDF*. Se pretende que, para cualquier

conjunto de datos o vocabulario, se ofrezca información relativa a la información que representa.

De esta forma, si una aplicación desea obtener información sobre un concepto identificado mediante una *URI*, cuando hace una llamada http para obtener el recurso, debería obtener información fácilmente procesable en formato *RDF*. De la misma forma, si se proveen puntos de consulta avanzada, como *SPARQL*, el resultado ante una consulta podrá se interpretado de forma automática (Harting, O. et al,. 2009).

La cuarta regla, enlazar datos en cualquier lugar, es necesaria para conectar los datos que tenemos en sitio web de forma que no se queden aislados y así se pueda compartir información con toras fuentes externas y que otros sitios puedan enlazar los datos propios de la misma forma que se hace con los enlaces en *HTML*.

A través del uso de enlaces a recursos provenientes de sitios más especializados en determinados dominios, se ofrece un valor añadido a la información que se provee.

Los enlaces de los recursos mediante *URIs*, pueden hacerse localmente y a través de toda la red, gracias a eso cualquier recurso es susceptible de ser enriquecido con cualquier tipo de información especializada, incluso la que no se espera que esté relacionada, de la misma forma al publicar información en *RDF* y utilizando *URIs*, cualquiera podría hacer referencia a esos datos (Doan, A. Et al,. 2012).

2.4 Modelo *RDF*

Resource Description Framework[11] (*RDF)* es un modelo estándar para el intercambio de datos en la Web, el cual tiene características que facilitan la fusión entre diferentes esquemas, y tiene soporte para la evolución de esquemas sobre el tiempo sin que se requiera que todos los datos del consumidor sean cambiados.

También extiende la estructura de vinculación de la Web para usar *URIs* en los nombres de las relaciones entre los objetos, que sirve para relacionar dos objetos y formar unta tripleta. Usando este modelo, es posible estructurar y semi-estructurar datos para ser mezclados, expuestos o compartido desde diferentes aplicaciones.

La estructura de vinculación es de forma directa, empleando un grafo etiquetado donde los bordes representas los links nombrados entre dos recursos, representado por los grafos de un nodo. Donde cada nodo esta conformado por tres elementos:

- **Sujeto**.- Puede ser el nodo inicial, una instancia, una entidad o una característica.
- **Predicado**.- Puede ser un verbo, una propiedad, un atributo, una relación, un miembro, un enlace o una referencia.
- **Objeto**.- Puede ser un valor, un nodo final o algún valor no literal que pueda ser usado como un sujeto.

Estos tres valores conforman una tripleta en RDF, como se ve en la Figura 2.3 y puede ser representada mediante una URI (*Uniform Resource Identifier*) o identificador de recursos únicos. Los sujetos y objetos son llamados nodos y pueden ser representados como un nodo en blanco. Los objetos pueden también ser representado como un valor literal. También un mismo nodo puede tener el rol de sujeto en una arista, y en otras ser un objeto.

[11] http://www.w3.org/RDF/

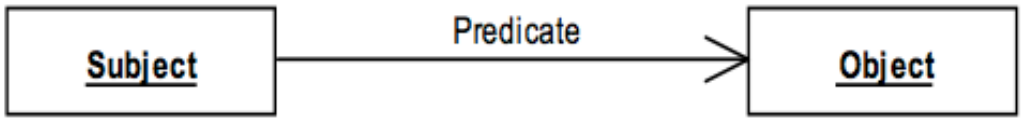


Figura. 2.3. Tripleta en RDF

2.5 Lenguaje de consulta SPARQL

SPARQL Protocol and RDF Query Language[12] *(SPARQL)* es el lenguaje de consultas para *RDF,* así como un protocolo con especificaciones para realizar consultas remotas.

*SPARQL Language*s un lenguaje, que es utilizado para realizar consultas a grafos en *RDF* a través de patrones de coincidencias. Este lenguaje permite incluir patrones básicos conjuntivos, filtros por valores, patrones opcionales y patrones de disyunción.

El protocolo *SPARQL* contiene una interfaz gráfica llamada *Sparql Query*, que permite realizar consultas. Para realizar consultas al protocolo SPARQL se necesita implementar un *Endpoint*, este es un servicio del protocolo *SPARQL* el cual esta definido conforme al estándar *SPROT*. Que permite a usuario realizar consultas a una base de conocimiento empleando el lenguaje *SPARQL*. La respuesta, puede ser retornada en diferentes formatos.

Tanto las consultas, como la presentación de resultados, debe ser implementada y recuperada en alguna aplicación independiente para poder ser interpretadas por un usuario, estas consultas utilizan como base los archivos RDF que se encuentran almacenados en un *Endpoint*, conocido también como *triple store.*

Un *triple store* es una base de datos especialmente diseñada para el almacenamiento y recuperación de tripletas, siendo una tripleta una entidad de

[12] http://www.w3.org/TR/rdf-sparql-query/

datos compuesta de sujeto-predicado-objeto, como "Luis tiene 30" o "Luis conoce Miguel". Al igual que una base de datos relacional, se almacena la información en un *triple store* y lo recupera a través de un lenguaje de consulta.

A diferencia de una base de datos relacional, *un triple store* está optimizado para el almacenamiento y recuperación de tripletas. Además de las consultas, las tripletas usualmente se pueden importar o exportar utilizando el formato *RDF* o en algún otro formato como *HTML* o *JSON* (Candillier et al., 2007).

2.6 GeoSPARQL

GeoSPARQL[13] es un estándar de la OGC, el cual define una extensión de funciones para *SPARQL* [W3r], un conjunto de reglas *RIF*[14] y un núcleo *RDF/OWL*[15] vocabulario para información geográfica basada en el *General Feature Model* (ISO 19109), *Simple Features* [ISO 19125-1], *Feature Geometry* (ISO 19107) y SQL MM (ISO/IEC 13249-3).

El estándar *GeoSPARQL* representa soporte y consultas de datos geoespaciales en la Web semántica. Éste define un vocabulario para representar datos en RDF, y una extensión para el lenguaje de consulta *SPARQL* para procesar datos geoespaciales.

El estándar sigue un diseño modular, que comprende diferentes componentes, los cuales se enumeran y describen a continuación:

1. **Un *core.-*** Define una clase *top-level* de *RDFS/OWL* para objetos espaciales.
2. **Un vocabulario topológico.-** Define propiedades *RDF* para afirmar y consultar las relaciones topologías entre objetos espaciales.

[13] http://www.opengeospatial.org/standards/geosparql
[14] http://www.w3.org/TR/rif-core/
[15] http://www.w3.org/RDF/

3. **Una geometría.-** Define tipos de datos *RDFS* para serializar datos geométricos, propiedades *RDF* geométricamente relacionadas y funciones topológicas no espaciales para objetos geométricos.
4. **Una topología geométrica.-** Define funciones topológicas de consulta.
5. **Una vinculación *RDFS.-*** Define un mecanismo de coincidencias implícitas en tripletas *RDF* que son derivadas en un *RDF* o en un esquema *RDFS*.
6. **Un *query rewriter***.- Define reglas, para transformar un patrón de tripletas a una relación topológica entre dos objetos en una *query* equivalente, que envuelve concretamente funciones de consulta geométrica y topológica.

Cada una de estas componentes forman parte de los requerimientos para *GeoSPARQL*, el cual está diseñado para integrarse a sistemas cualitativos basados en razonamiento espacial y sistemas cuantitativos basados en computo espacial.

2.7 Big Data

El *Big Data* es una nueva tendencia científica [16] Impulsada por el análisis de datos en alta dimensión, la tecnología de big data elabora correlaciones de datos (indicadas por parámetros estadísticos) para obtener información sobre los mecanismos inherentes[17] . Los resultados basados en datos solo se basan en una selección desenfrenada de datos sin procesar del sistema (el espacio puede ser un sistema completo o solo una región, el tiempo puede ser largo o corto, y el tamaño puede ser grande o pequeño) y un procedimiento estadístico general (para procesamiento de datos). Por otro lado, los procedimientos para el análisis tradicional basado en modelos, especialmente el desacoplamiento de un sistema práctico interconectado, siempre se basan en suposiciones y simplificaciones. Los resultados basados en modelos se basan en causalidades identificadas, parámetros específicos, selecciones de muestras y procesos de capacitación; las fórmulas / expresiones imprecisas o incompletas, las selecciones de muestra sesgadas y los

[16] http://www.nature.com/news/specials/bigdata/index.html
[17] http://www.sciencemag.org/site/special/data/.

procesos de capacitación inadecuados conducirán a malos resultados. Los resultados son a menudo apenas satisfecho o incluso insatisfecho a medida que crece el tamaño del sistema y aumenta la complejidad. En términos generales, las herramientas de análisis impulsadas por datos, en lugar de las basadas en modelos, son más adecuadas para complejos sistemas interconectados a gran escala con datos fácilmente accesibles.

La tecnología de Big Data no entra en conflicto con los análisis o pre tratamientos clásicos. En realidad, la tecnología de Big Data ya se ha aplicado con éxito como una poderosa herramienta basada en datos para numerosos fenómenos, como sistemas cuánticos (Brody, T. A., et al,. 1981), sistemas financieros (Laloux, L., et al,. 2000) (Chen, H., et al,. 2012), sistemas biológicos (Howe, D., et al,. 2008) , así como redes de comunicación inalámbricas (Qiu, R., & Wicks, M., 2014) y (Li, X., Lin, F., & Qiu, R. C., 2014).. Las principales tareas de la arquitectura para estas aplicaciones parecen similares: 1) el modelado de Big data y 2) el análisis de Big Data. Se cree que la tecnología de Big data también tendrá un amplio alcance aplicado en los sistemas de potencia, y los resultados serán fructíferos.

2.8 Herramientas para BIGDATA

Actualmente existen varios framework que permiten realizar un análisis de datos con Big Data, o herramientas que facilitan el uso de las técnicas de minería de datos. Entre los frameworks más usados se encuentran los desarrollados por Apache: Hadoop y Spark, los cuales permiten el procesamiento de los datos a gran escala. Para desarrollar herramientas personalizadas, es decir, enfocadas a una temática en especial, o la implementación de un software propio, se puede hacer uso de los entornos de desarrollo de programación, con esto se realiza aplicaciones más robustas, enfocadas al área de aplicación. A continuación, se describe los frameworks y herramientas más usados en analítica de datos.

2.8.1 Apache Hadoop

Apache Hadoop es un framework que permite el procesamiento distribuido de grandes conjuntos de datos usando simples modelos de programación. Está diseñado para pasar de simples servidores a miles de máquinas computacionales con almacenamiento propio. En lugar de depender de hardware para ofrecer alta disponibilidad, se ha diseñado para manejar fallas en la capa de aplicación (Hadoop, A, 2014). HDFS, de sus siglas en inglés Hadoop distributed file system, es un sistema de archivos distribuido, y está compuesto de un único NameNode que es un servidor maestro. Administra el espacio de nombres del sistema de archivos y está vinculado a un Datanode para administrar el almacenamiento de datos. Esta estructura implica que todos los bloques de un archivo se pueden 15 guardar en varias máquinas. La información de grandes cantidades de datos (metadatos) consiste en particiones de archivos en bloques y la distribución de estos bloques en diferentes Datanodes. En la Figura 2.4, se muestra la interfaz web de Apache Hadoop (Hadoop, A, 2014).

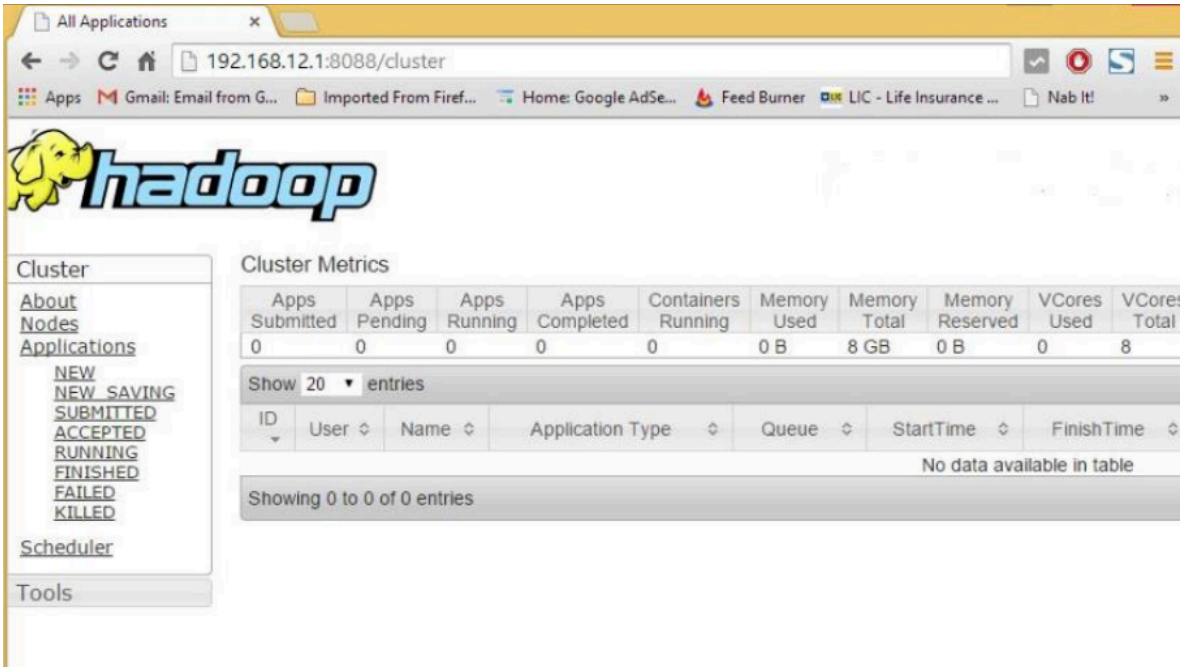


Figura 2.4. Logo de Apache Hadoop Fuente: (Hadoop, A, 2014) 2.3.1.2.

2.8.2 Apache Spark

Spark es un framework considerado como un motor de procesamiento (denominado DAG) que se ha construido con base en la velocidad, facilidad de uso y análisis sofisticados. Permite la gestión de datos de diferente naturaleza como son los textos o los gráficos, además puede procesar a gran escala usando petabytes de datos en múltiples clusters con un mayor número de nodos. El acceso a la intefaz web del framework Apache Spark se muestra en la Figura 2.5. (Apache Spark, 2016).

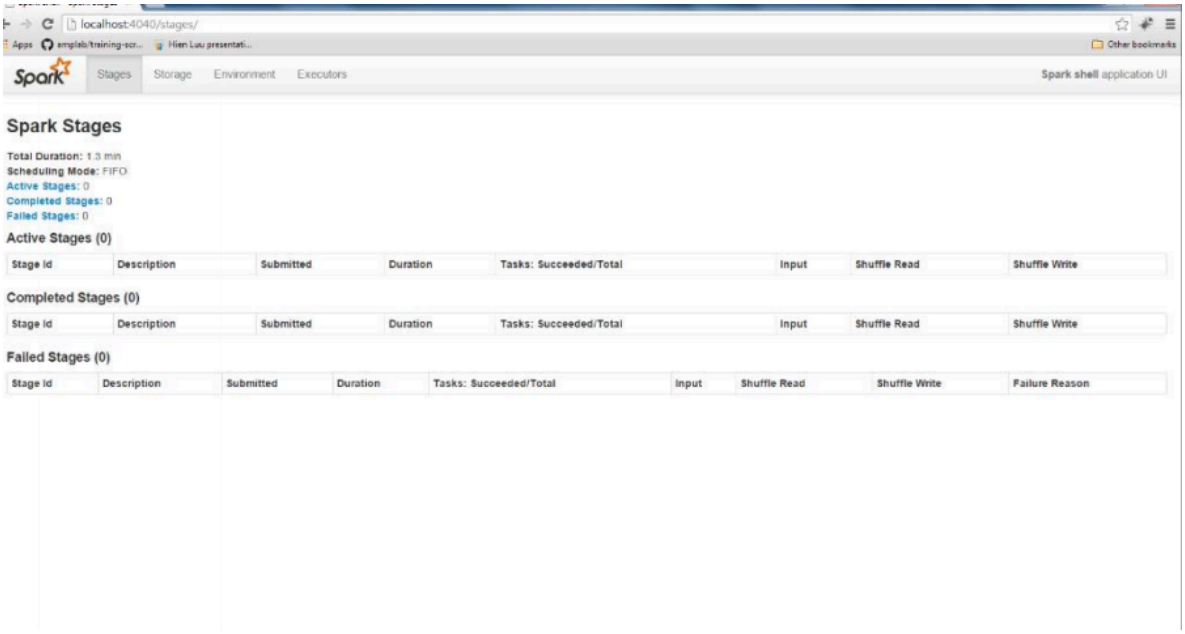


Figura 2.5 Uso de la interfaz web de Spark. Fuente: (InfoQ, 2014).

2.8.3 Apache Kafka

Apache Kafka® es una plataforma de transmisión distribuida. Una plataforma de transmisión tiene tres capacidades clave:

1) Publique y suscríbase a secuencias de registros, de forma similar a una cola de mensajes o un sistema de mensajería empresarial.

2) Almacene flujos de registros de una manera duradera y tolerante a fallas.

3) Procesar flujos de registros a medida que ocurren.

Kafka generalmente se usa para dos amplias clases de aplicaciones:

1) Construye canales de datos de transmisión en tiempo real que obtienen datos confiablemente entre sistemas o aplicaciones.
2) Construye aplicaciones de transmisión en tiempo real que transforman o reaccionan a las corrientes de datos.

Para entender cómo Kafka hace estas cosas, profundicemos en algunos conceptos:

Kafka se ejecuta como un clúster en uno o más servidores que pueden abarcar múltiples centros de datos. El clúster de Kafka almacena secuencias de registros en categorías llamadas temas. Cada registro consta de una clave, un valor y una marca de tiempo.

Kafka tiene cuatro API principales Figura 2.6:

1. Producer API: permite que una aplicación publique una secuencia de registros de uno o más temas de Kafka.
2. La API de consumidor permite a una aplicación suscribirse a uno o más temas y procesar la secuencia de registros que se les produce.
3. La API de *Streams* permite que una aplicación actúe como un procesador de flujo, consumiendo un flujo de entrada de uno o más temas y produciendo un flujo de salida para uno o más temas de salida, transformando de manera efectiva los flujos de entrada a los flujos de salida.
4. Connector API permite construir y ejecutar productores o consumidores reutilizables que conectan temas de Kafka con aplicaciones o sistemas de datos existentes. Por ejemplo, un conector a una base de datos relacional puede capturar cada cambio en una tabla.

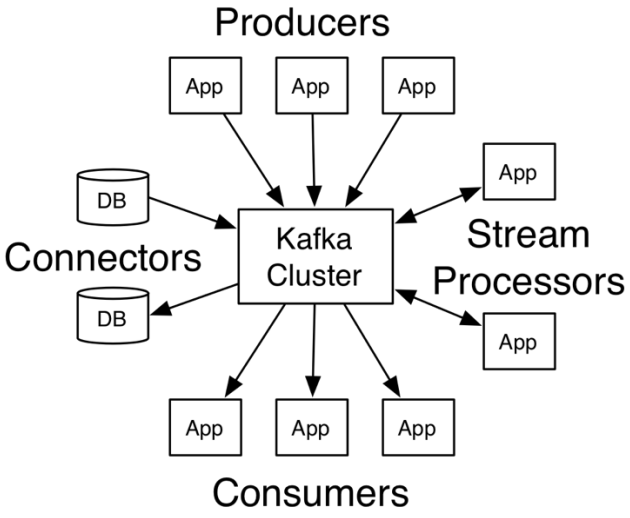


Figura 2.6 Arquitectura de Apache Kafka.

En Kafka, la comunicación entre los clientes y los servidores se realiza con un protocolo TCP simple, de alto rendimiento, independiente del idioma. Este protocolo está versionado y mantiene la compatibilidad con versiones anteriores (Apache Kafka, 2016).

CAPÍTULO 3. ESTADO DEL ARTE

En este capítulo se describen algunos trabajos relacionados con integración semántica, ingeniería ontológica, Big Data, información geográfica voluntaria (VGI) entre otros.

3.1 Trabajos relacionados con Big Data

Existe una creencia generalizada de que Big Data puede ayudar a mejorar las predicciones siempre que podamos analizar y descubrir patrones ocultos, y en (Richards y King, 2013) mencionan que las predicciones pueden mejorarse a través de la toma de decisiones basada en datos. (Tucker, 2013) cree que Big Data pronosticará pronto todos nuestros movimientos, y de acuerdo con (Einav y Levin, 2014), Big Data es el más buscado para construir modelos predictivos en un mundo en el que las previsiones continúan siendo un problema estadístico vital (Hand, 2009). (Hassani y Silva, 2015) llegan a la pregunta, ¿cuál es el problema detrás de la previsión con Big Data? Mencionan que la explicación más simple es que las herramientas de pronóstico tradicionales no pueden manejar el tamaño, la velocidad y la complejidad inherentes a Big Data (Madden, 2012). Según (Arribas-Bel, 2014) esto se debe a la falta de una estructura en estos conjuntos de datos y al tamaño. Como resultado, rara vez se prefieren las técnicas tradicionales para abordar Big Data (Arribas-Bel, 2014). Por lo tanto, pronosticar Big Data representa un desafío para las organizaciones.

(Rey y Wells, 2012) creen que las técnicas de Minería de Datos pueden explotarse para ayudar a pronosticar con Big Data, una opinión respaldada por (Varian, 2014). Sin embargo, debe tenerse en cuenta que, en el pasado, las técnicas de minería de datos se han utilizado principalmente en datos estáticos en oposición a series temporales (véase, por ejemplo (Ghodsi, 2014), (Pyke, 2003); (Kurgan y Musilek, 2006); (Han y Kamber, 2011).

Las oportunidades de obtener beneficios a través de la previsión con Big Data son diversas. En la actualidad, hay una mayor investigación sobre el uso de Big Data para obtener pronósticos meteorológicos precisos y los resultados iniciales sugieren que Big Data beneficiará inmensamente a los pronósticos meteorológicos (Hamm, 2013), (Knapp, 2013). De hecho, la previsión meteorológica ha sido uno de los principales beneficiarios de Big Data, pero las previsiones siguen siendo inexactas más allá de una semana (Silver, 2012). De acuerdo con (Bacon, 2013), la industria de las aerolíneas es otro campo donde la previsión de Big Data es crucial. Una historia de éxito interesante detrás de la previsión con Big Data es Netflix y su uso de pronósticos de Big Data para la toma de decisiones antes de comenzar la producción de su propio programa de televisión "House of Cards", y esto dio como resultado mayores ingresos para la compañía. El potencial subyacente a los pronósticos de Big Data es realmente asombroso y a veces 'aterrador' como fue evidente en la experiencia de un individuo en una historia narrada por (Duhigg, 2012) donde un cliente iracundo entra a una tienda 'Target' en Minneapolis para quejarse de la tienda enviando cupones relacionados con productos de embarazo a su hija de secundaria. Unas semanas más tarde, el mismo cliente se disculpa con el gerente, ya que después de una discusión con su hija, se reveló que, de hecho, estaba embarazada (Duhigg, 2012).

En (Brende, et al., 2015) presentan una idea sobre cómo descubrir conocimientos adicionales de datos de agricultura de precisión a través del enfoque de big data. Los autores presentan un escenario para el uso de los servicios de tecnología de la información y la comunicación (TIC) en el entorno de big data agrícola para recopilar grandes cantidades de datos. El análisis de Big Data en aplicaciones agrícolas brinda una nueva perspectiva para tomar decisiones meteorológicas anticipadas, mejorar la productividad del rendimiento y evitar costos innecesarios relacionados con la cosecha, el uso de pesticidas y fertilizantes. Utilizan un modelo de programación y un algoritmo distribuido para el procesamiento de datos y la aplicación de predicción del clima.

Combinar cantidades cada vez mayores de energía renovable en redes de suministro eléctrico requiere estimaciones precisas de la potencia de esos recursos para la planificación diaria y las operaciones en tiempo real. Esto requiere predecir el recurso eólico y solar en esas escalas de tiempo. La predicción precisa de estas variables meteorológicas es un problema de big data que requiere una gran cantidad de datos dispares, múltiples modelos que son aplicables a un marco de tiempo específico y la aplicación de técnicas de inteligencia computacional para combinar con éxito todo el modelo y la información observacional en tiempo y entregarlo a los tomadores de decisiones en los servicios públicos y los operadores de la red. Teniendo en cuenta que la capacidad de las energías renovables continúa creciendo, un reto adicional incluye seleccionar y archivar datos para el reciclaje continuo de los algoritmos de aprendizaje automático (Haupt y Kosovic, 2015).

3.2 Ingeniería ontológica

En (Cruz et al., 2017) proponen el uso de sistemas multi-agente y ontologías para solucionar el problema de monitoreo de procesos heterogéneos cuyos datos son adquiridos mediante sensores. Los autores mencionan que la representación de información mediante ontologías permite un mejor control de proceso, considerando que el monitoreo es un aspecto fundamental en la automatización de procesos industriales. Para probar el diseño de la ontología, la implementaron para solucionar el problema de controlar un proceso industrial, simulando el proceso mediante un sistema multi-agente.

El trabajo presentado por (Morales et al., 2017) presenta un sistema ontológico como herramienta en un dominio o subdominio especifico, relaciona o interactúa con conjunto de partes armonizadas de un conocimiento real. Desarrollaron un prototipo ontológico, que permite monitorear ecosistemas dulceacuícolas mediante el componente biótico de una manera rápida y eficiente en la valoración de la calidad de agua.

En (Ramos & Núñez, 2007) desarrollan una ontología que consiste en una aplicación en ambiente web para la visualización de la ontología de insectos acuáticos, el diseño permite acceder a este conocimiento de manera sencilla sin límites espaciales en cualquier ubicación geográfica; en ella se presenta una aplicación de visualización para ambiente Web, la cual le ofrece a los usuarios funcionalidades como: navegar en la Ontología y consultar el contenido de la misma de forma gráfica y textual, visualizar imágenes asociadas a instancias específicas, descargar documentos, y acceder a sitios de interés asociados con el dominio. Por otra parte, desde el punto de vista ambiental, en la evaluación de impactos ambientales (EIA), se presenta el trabajo de Garrido (2011), en el que propone una utilización ontológica que garantiza utilizar mejores técnicas de esta rama del conocimiento, dada la complejidad de las EIA. De igual manera, se presenta una ontología para proyectos mineros (Montes de Oca-Pérez y Rosario-Ferrer, 2004, Moreno et al., 2013). En los sistemas geoespaciales, las otologías también han tenido sus acogidas formales en este campo. En este sentido, se tienen los trabajos de Blázquez et al., (2008); Larin-Fonseca y Garea-Llano (2013); Vilches-Blázquez y Ramos (2012), en el que se expone los diferentes enfoques existentes para llevar a cabo la conflación semántica entre diversos conjuntos de datos geoespaciales, información geográfica y la definición de los significados de los fenómenos coherentes de la realidad.

3.4 Trabajos relacionados con VGI

VGI es un término acuñado por Goodchild (Goodchild, 2017), utilizado para definir el uso de la web con el objetivo de crear, recopilar y difundir datos geográficos proporcionados voluntariamente por ciudadanos comunes. Aunque VGI es el término más ampliamente utilizado, existe una discusión sobre si la palabra “voluntario” es apropiada o no (Gorman, 2012): no todos los datos de VGI son proporcionados solo por los usuarios, debido a que algunas organizaciones pueden ayudar en la tarea de obtener información. Además, el voluntariado no siempre es el adjetivo adecuado, ya que algunos sistemas de información geográfica

colaborativos (Harvey, 2013), el usuario debe conocer toda la información que proporciona y decidir cuándo ofrecerla para que sea voluntaria, y no todos se darán cuenta de que su teléfono inteligente está enviando datos de ubicación a Internet.

En los últimos años, el mercado de teléfonos inteligentes ha alcanzado niveles particularmente altos; este hecho hace que el proceso de compartir información de fuentes multitudinarias (crowdsourcing) (voluntariamente o no) sea más fácil y más rápido. El nivel de precisión de los datos está aumentando de acuerdo con las mejoras en los dispositivos en cuanto a software y hardware, pero todavía hay algunos problemas con los usuarios de GIS móviles con respecto a la conexión a Internet porque no siempre es posible trabajar conectado debido a conexiones ruidosas, limitadas o inseguras (Longley, 2011).

VGI ha tenido un gran impacto en las Ciencias de Información Geográfica, porque la recopilación de datos solía ser una de las tareas que requería más tiempo y esfuerzo. Hace algún tiempo, todos los datos geográficos debían recopilarse manualmente o ser localizados por expertos en Sistemas de Información Geográfica, hoy en día todo el mundo puede actuar como creador de datos geográficos, ya que cada ciudadano genera una gran cantidad de datos cuando utiliza sus teléfonos inteligentes.

Además, la cantidad de sensores integrados en los dispositivos también ha aumentado drásticamente. La combinación de estas dos tendencias ha permitido un nuevo tipo de investigación llamada detección enfocada en las personas (Campbell, et al. 2008, Miluzzo et al. 2008, Lane et al. 2010). Por lo tanto, los sensores en los teléfonos móviles y otros dispositivos inalámbricos se pueden utilizar para recopilar grandes cantidades de mediciones continuas sobre los usuarios. Las diferentes modalidades de datos recopilados a través del cliente se pueden clasificar de la siguiente manera: datos de interacción social, que pueden inferirse a partir de registros de llamadas, registros de mensajes cortos, resultados de búsqueda de dispositivos Bluetooth, muestras del entorno acústico, etc.; datos

de ubicación, los cuales se pueden determinar en función del GPS (cuando este habilitado), la información de la red celular y la información del punto de acceso WLAN (cuando esté disponible); Creación de medios y datos de uso: se puede capturar información sobre lugares donde se han capturado imágenes, grabación de videos y donde se ha reproducido música; y, datos de comportamiento: se puede capturar información sobre el uso de una aplicación, la detección de la actividad basada en el sensor de aceleración y las estadísticas de uso de dispositivos regulares basados en registros de llamadas y mensajes cortos.

3.5 Censado de fuentes multitudinarias (Crowd Sourced Sensing)

En el sensado de fuentes multitudinarias, un grupo de usuarios profesionales o privados recolectan y contribuyen la información de forma colaborativa para formar un cuerpo de conocimiento. Particularmente, el aumento de telefonos inteligentes y “la creciente capacidad de capturar, clasificar y transmitir una amplia variedad de datos (imagen, audio y ubicacion) ha habilitado un nuevo paradigma de sensado” (Reddy et al., 2007). Las agencias cívicas de varios países alrededor del mundo ya están aprovechando acelerando y ampliando el uso de métodos de innovación abiertos para ayudar a conducir hacia una amplia gama de problemas urbanos y sociales que van desde la observación de la vida silvestre hasta el monitoreo de la calidad del aire. El geo sensado colaborativo se ha aplicado con éxito para estudiar los fenómenos físicos, particularmente en contextos urbanos, tales como monitoreo de niveles de ruido urbano como alternativa al monitoreo ambiental tradicional (D’Hondt et al., 2013).

En el internet de las cosas cualquier cosa se puede medir utilizando un conjunto de observaciones que reducen la incertidumbre cuando el resultado se expresa como una cantidad (Hubbard, 2007). Este punto de vista estadísticamente motivado sobre la medición contradice parcialmente la visión clásica sobre los procesos de medición tan estandarizados como DIN 1319 el cual se ha moldeado desde el siglo pasado. Teniendo en cuenta la escasa solución espacial y temporal

de muchas mediciones disponibles en la actualidad, cualquier cosa es mejor que adivinar (además del conocimiento existente) potencialmente puede contribuir a una medición, incluso si no se considera una medida por definiciones estrictas. Sin embargo, esto requiere algoritmos para abordar el "problema de la interrelación entre la fiabilidad de las fuentes de información, su número y la fiabilidad de los resultados de la integracion" (Rogova et al., 2004).

Las primeras investigaciones, centradas principalmente en la administración de sensores distribuidos en redes de sensores como la de IrisNet de Intel (Gibbons et al., 2003) o la de SenseWeb de Microsoft (Grosky et al., 2007), se han hecho realidad desde hace varios años debido a la amplia variedad de dispositivos como NetAtmo, y han atraído la atención de investigadores, particularmente interesados en datos de mayor resolución (Chapman et al., 2017), (Meier et al., 2017).

3.6 Predicción de factores ambientales

La principal ventaja de VGI es obtener más datos e información sobre el entorno para formular y evaluar hipótesis y obtener información valiosa sobre el medio ambiente. Los datos obtenidos se utilizan para entrenar modelos para predecir factores ambientales como la temperatura y la contaminación. La base para todos los modelos de predicción especial es la Primera Ley de Tobler (Tobler, 1970), afirma que "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes". Uno de los enfoques más utilizados para incorporar esta ley es el de Kriging (Krige, 1951), el cual se desarrolló para estimar los depósitos de minerales, pero desde entonces se ha utilizado para realizar predicciones de una variedad de aplicaciones espaciales y ha sido modificado para que sea más potente y general. Hengl et al. (2012) utiliza un enfoque de kriging para predecir las temperaturas. Incluyen una componente temporal para predecir la temperatura media diaria en Croacia para un área de 1 km^2 con una precisión de 2.4°C mediante la combinación de imágenes de satélite Modis con 57,282 mediciones terrestres de las temperaturas en 2008. En un trabajo de seguimiento,

Kilibarda et al. (2014) presentaron un marco para un mapeo automatizado para realizar predicciones de las temperaturas del aire diarias medias, mínimas y máximas utilizando una regresión kriging para una resolución de 1km con un error de raíz cuadrada media entre 2 y 4°C.

Gräler et al. (2016) desarrollaron un paquete en R llamado gstat, el cual utiliza copulas para habilitar un kriging espacio-temporal. Los autores muestran, además, el potencial de su enfoque con una predicción de la concentración media diaria de 10PPM en el 2005 en Alemania.

Otra modificación del enfoque kriging se puede encontrar en Bhattacharjee et al. (2016). El cual propone un enfoque kriging semántico, donde se utiliza una fotografía de satélite instantánea de alta resolución para cuantificar el efecto de la diferencia entre diferentes lugares, así como la interacción de los usos de la tierra en esos puntos. Las diferentes clases de uso del suelo se aprenden en una red semántica jerárquica.

Hjort et al (2011) presentaron un enfoque diferente para predecir las temperaturas locales en la ciudad de Turku, Finlandia. Los autores utilizaron modelos lineales generalizados combinados con árboles de regresión y datos de 36 estaciones meteorológicas estacionarias durante un periodo de seis años. Una visión general fundamental y antecedentes teóricos, asi como las aplicaciones de estadísticas espacio-temporales se pueden encontrar en Cressie and Wikle (2015).

Australia es propenso a los impactos devastadores de incendios forestales y otros peligros naturales, en (Haworth et al., 2015) los autores presentan los resultados de un estudio que examina el papel potencial de las redes sociales y otras tecnologías de información geográfica en línea para fomentar la participación de la comunidad en la preparación de incendios forestales en Tasmania. El cambio climático y el aumento del calentamiento global provocan que los eventos climáticos extremos, como los incendios forestales, las inundaciones y las olas de calor,

aumentarán tanto en frecuencia como en intensidad (IPCC 2012). Prepararse adecuadamente para los desastres puede reducir drásticamente el riesgo para la vida y los bienes (Paton 2003). Sin embargo, a pesar de los esfuerzos para educar a las comunidades con información relevante y actualizada, la investigación ha demostrado que las personas en comunidades en riesgo aún no pueden participar activamente en actividades de reducción de riesgos (Frandsen 2011). Se necesitan enfoques innovadores para involucrar a las comunidades en la preparación para desastres a fin de reducir los riesgos y crear resiliencia. Las redes sociales y otras tecnologías de comunicación de información geográfica en línea ofrecen cada vez más oportunidades para conectar a las comunidades. El papel de estas tecnologías en la respuesta a desastres ha sido bien establecido en los últimos años, sin embargo, la investigación sobre su utilidad en las fases previas al desastre del ciclo de emergencia sigue siendo relativamente limitada (Haworth y Bruce 2015).

Los medios sociales son aplicaciones basadas en Internet que permiten a las personas comunicarse y compartir recursos (Taylor et al., 2012). Otras tecnologías de comunicación de información geográfica a las que se hace referencia en este artículo incluyen software de creación de mapas en línea abierto a contribuciones públicas (por ejemplo, Ushahidi Crowdmap, OpenStreetMap) y dispositivos como teléfonos inteligentes, que permiten recopilar, crear e intercambiar datos de formas sin precedentes. El compromiso generalizado del público para producir voluntariamente información geográfica utilizando estas tecnologías se conoce como información geográfica voluntaria (VGI) (Goodchild 2007). Antes de la aparición de VGI, la información geográfica de la comunidad se recopiló mediante grupos focales, encuestas y debates comunitarios, y se demostró que el conocimiento local, tradicional e indígena es útil tanto en la gestión ambiental como en el mapeo de desastres (Prober et al., 2011, Tran et al. al. 2009). A pesar de los desafíos significativos, particularmente los de calidad, precisión y credibilidad de los datos (ver Flanagin & Metzger 2008, Elwood, Goodchild & Sui 2012), VGI en la gestión de desastres permite una recopilación y difusión rápida y económica de información local diversa, con grandes cantidades de datos recolectados casi en

tiempo real. Permite una mayor conectividad con las comunidades y las autoridades y facilita la comprensión del riesgo local a través del mapeo y el intercambio de conocimiento local.

El cambio climático es un fenómeno espacio-temporal y una cuestión mundial también. Desafía el medio ambiente sostenible y el desarrollo. Se prevé que el cambio climático afectará a casi todos los sectores importantes, como la agricultura y la seguridad alimentaria, el agua y los recursos energéticos, la infraestructura física, los bosques, la biodiversidad y el medio ambiente marino (Ahmada et al., 2016). Con el fin de desarrollar estrategias y planes de acción para la adaptación y mitigación de los efectos del cambio climático, se requiere información espacio-temporal de varios tipos actualmente bloqueada por varias organizaciones que no dan acceso a sus conjuntos de datos en países como Pakistán.

En (Ježek et al., 2015) presentan un proceso de extracción de partes relevantes de datos abiertos y fuentes de datos VGI que son adecuados para la predicción y el análisis del historial de tráfico. En particular, se enfocan en la predicción del volumen de tráfico (la cantidad de vehículos que pasan por un segmento de red).

La información geográfica voluntaria (VGI) se ha utilizado para complementar o sustituir información autorizada en el dominio de la gestión de inundaciones. El principal problema con respecto al uso de la información voluntaria es estimar su calidad, principalmente porque puede sufrir de una calidad heterogénea. Por lo tanto, se han desarrollado varios métodos en los últimos años para evaluar la calidad de VGI. Sin embargo, las obras existentes no tienen en cuenta la calidad de VGI para el manejo de inundaciones. Para superar esta brecha, (Castro Degrossi et al., 2017) proponen un método para evaluar la calidad de VGI para este fin. Este método usa un conjunto de métricas de calidad que fueron desarrolladas para medir la plausibilidad de VGI. Realizan una regresión lineal múltiple para demostrar la relación entre la verosimilitud de VGI y las métricas de calidad. Los resultados

mostraron que la verosimilitud se puede explicar mediante 5 métricas de calidad. Por lo tanto, el método propuesto es capaz de estimar la verosimilitud de VGI en el dominio de gestión de inundaciones.

En (Castro Degrossi et al., 2017) proponen un método para evaluar la calidad de VGI, principalmente porque puede obtenerse de fuentes heterogéneas. El método que proponen los autores es un conjunto de métricas que fueron desarrolladas para medir la plausibilidad de VGI. Aplicaron su método para estimar la verosimilitud de VIG en el dominio del control de inundaciones.

En los últimos años, la información geográfica voluntaria ha ganado especial atención en el dominio de la gestión de inundaciones debido a su potencial para optimizar la detección de eventos de inundaciones (Longueville et al., 2010; Ludwig et al., 2015), sirven también para la toma de decisiones (Ahmad y Simonovic 2006; Horita et al., 2015; Simonovic 1999), mejoran el pronóstico de las inundaciones (Mazzoleni et al., 2017) y complementan los datos autorizados (Degrossi et al., 2014; Lanfranchi et al., 2014). Potencialmente, la información geográfica puesta a disposición por los voluntarios puede contribuir a minimizar la incertidumbre en la predicción de inundaciones y también a mejorar la respuesta a los eventos de inundación (Ostermann y Spinsanti 2011).

Considerando el dominio de la prevención de inundaciones, Poser y Dransch (2010) han utilizado datos fidedignos para estimar la calidad de VGI. Los autores compararon la profundidad de inundación proporcionada por los voluntarios con la medida por una fuente autorizada. Del mismo modo, Moreira et al. (2015) y Degrossi et al. (2014) han estimado la calidad de las observaciones del nivel del agua al compararlas con datos del sensor en tiempo real. Sin embargo, este método a menudo está limitado por la disponibilidad de datos fidedignos (Hung et al., 2016) y los costos y las restricciones de licenciamiento (Mooney et al., 2010). Una solución alternativa es analizar el contexto geográfico. Hung et al. (2016) analizaron factores

de geolocalización para evaluar la credibilidad de VGI en un escenario de respuesta a inundaciones.

CAPÍTULO 4 METODOLOGÍA

En este capítulo se describe, de manera general, la metodología propuesta, y de manera exhaustiva las etapas que componen la misma describiendo las técnicas, estándares y protocolos implementados en cada una de las etapas.

4.1 Descripción de la metodología

La metodología propuesta en este trabajo está enfocada a la integración de diversas y heterogéneas fuentes de datos meteorológicos y calidad del aire (estaciones base meteorológicas, servicios de monitoreo con *APIs* o servicios web), procedentes de fuentes oficiales y voluntarias.

Según (Vilches-Blázquez, 2011) la integración semántica permite mezclar las respuestas recuperadas con las bases de datos públicas y/o otras fuentes de información disponibles, de tal forma que la información pública debe ser accesibles en diversos formatos. Acorde a esto se puede relacionar directamente la integración con Big Data de la siguiente manera:

Volumen.- Este componente se ve representado por la cantidad de datos recuperados de manera dinámica y los procesados de manera estática, teniendo alrededor de 24 a 48 medidas por estación base registrada.

Variedad.- Este componente se ve en la manera en que la información recuperada esta representada en diferentes formatos concretamente JSON y XML de la parte de fuentes dinámicas y CSV del lado de las oficiales.

Velocidad.- Este componente se ve reflejado en las constantes mediciones que se realizan en las estaciones meteorológicas y de calidad del aire tanto de las fuentes voluntarias (tiempos variables según el proveedor) como de las oficiales.

Veracidad.- Este componente se refleja en la calidad de los datos recuperados, por el lado de las fuentes voluntarias la calidad de los datos es variable por diferentes motivos, mientras que por las fuentes oficiales lo datos son de gran calidad.

Valor.- Este componente precisamente en la integración de los dos tipos de fuentes de información, esto debido a que las fuentes oficiales pueden ser complementadas con las fuentes voluntarias, esto deriva en un mayor valor de los datos integrados.

Por lo tanto, el proceso integración semántica esta directamente relacionado con la interoperabilidad semántica, que se refiere a los mecanismos que permiten a dos o mas sistemas compartir e integrar información desde diferentes fuentes con el fin de superar los problemas que provoca en la información la heterogeneidad semántica (Vilches-Blázquez, 2011).

En el caso de nuestra metodología utilizaremos una red de ontologías, la cual se compone de varias ontologías usadas para integrar semánticamente las diversas mediciones meteorológicas y de calidad de aire recuperadas de las fuentes de información, además de emplear los principios de *Linked Data* que sirven para compartir información por medio de la vinculación a otras fuentes de información, esto conlleva al uso de puntos de consultas avanzadas *SPARQL* para recuperar más información toda en formato RDF.

Tras la integración semántica, se obtendrá como resultado un repositorio de observaciones, asociadas con diversas variables meteorológicas y de calidad del aire, semánticamente integradas que continuamente seguirá creciendo de manera constante con la información recuperada de las fuentes de datos voluntarias.

De manera más específica, la metodología desarrollada en este trabajo posee cuatro componentes como se ve en la Figura 4.1.: El primer componente ese centra en la recolección de datos, que se lleva a cabo de dos maneras diferentes:

estática y dinámica; El segundo componente es el asociado con el pre-procesamiento de los datos que han sido recuperados por el componente anterior; El tercer componente es el módulo de integración semántica. Este componente recoge el detalle de la implementación de la red ontológica, utilizando diferentes escenarios de la metodología NeOn (Suárez et al., 2012), la población de la red ontológica a través de un proceso automático de generación de RDF y el establecimiento de conexiones a la nube de *Linked Data* para enriquecer los datos recuperados. Finalmente, el componente de análisis que permite explotar los datos integrados semánticamente a través de diversas operaciones de análisis.

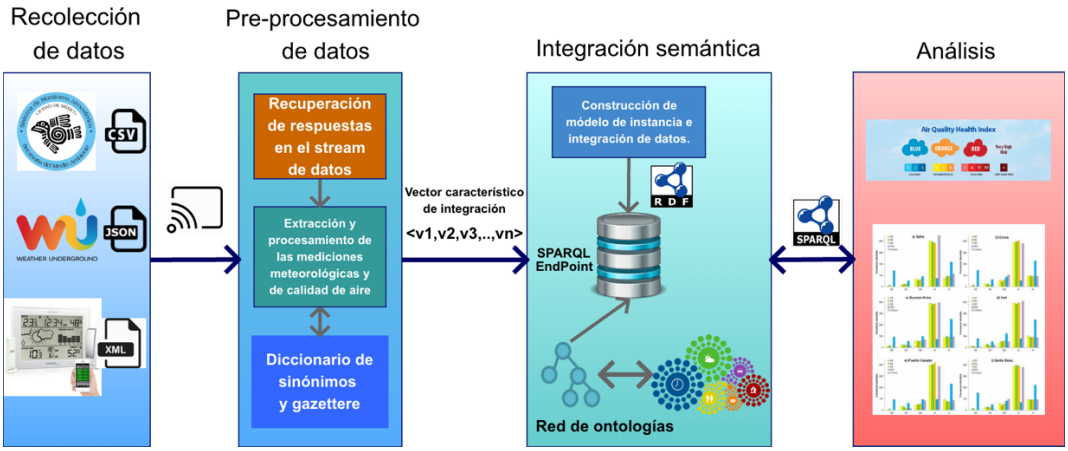


Figura 4.1. Diagrama general de la metodología propuesta

4.2 Recolección de datos

En este primer componente es donde se realiza el proceso de recuperación de información de fuentes de información oficiales y voluntarias como se ve en la Figura 4.2.

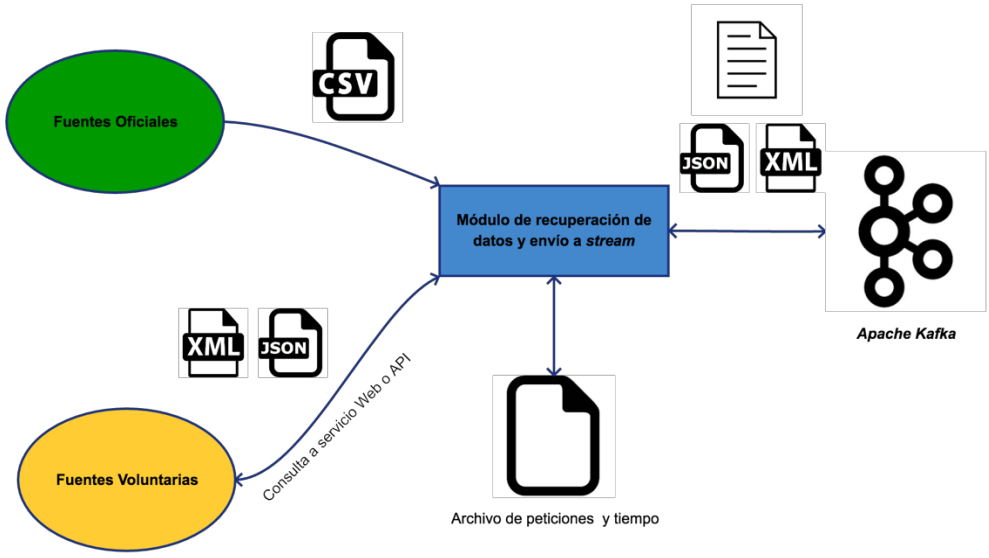


Figura 4.2. Proceso de recolección de datos

Este componente posee dos tipos de fuentes, dependiendo del tipo de fuente a tratar (estática o dinámica). Así, si las fuentes de información son estáticas se procede a la carga y lectura de archivos CSV, estas medidas se definen con la siguiente matriz:

$$CSV = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

Donde:

- **n**.- Sea el número de variables que contiene el archivo csv.
- **m**.- Sea el número de mediciones que contiene el archivo csv.

Esta matriz de entrada debe contar con el encabezado que contenga los nombres de las variables que se van a procesar, por lo tanto, debe cumplir con la siguiente condición:

$$CSV_{(1,1\ldots n)} = \{x | \; x \; es \; el \; nombre \; de \; cada \; variable \; del \; archivo\}$$

Por lo regular estas fuentes de información provienen de instituciones gubernamentales como por ejemplo la REDMET, RAMA o CONAGUA entre otras como se puede ver en la Tabla 4.1. Para las fuentes de información dinámica se procede a realizar peticiones a los servicios web y Apis disponibles como Aeris Weather, WeatherUndeground, entre otras, como se puede ver en la tabla 4.1 y recuperando sus respuestas, por lo regular en formato JSON o XML.

Tabla 4.1. Clasificación de fuentes heterogéneas disponibles

Fuente de datos	Tipo de fuente	Tipo de respuesta o archivos
CONAGUA	Oficial (MX)	CSV
REDMET	Oficial (MX)	CSV
SEMAR	Oficial (MX)	CSV
3TIER	Voluntaria	JSON
AccuWeather	Voluntaria	JSON
FAA Airport Service	Oficial (EUA)	XML o JSON
Aerisapi	Voluntaria	XML o JSON
Weatherbug	Voluntaria	JSON
WeatherUnderground	Voluntaria	JSON
GEONAME	Voluntaria	JSON o XML
Foreca	Voluntaria	JSON o XML

Para la recolección estática primero procedemos a realizar la lectura del archivo CSV utilizando el encabezado del archivó para saber que medidas son las contiene como se ve en la Tabla 4.2, posteriormente se transforma en un objeto de tipo texto plano en le cual se almacena cada renglón del archivo adjuntando de quien es el proveedor del archivo como se ve en la Figura 4.3.

Tabla 4.2. Ejemplo de archivo CSV

date	Hora	cve_station	cve_parameter	value
01/01/11	1	ACO	RH	
01/01/11	1	MON	RH	38
01/01/11	1	CHO	RH	
01/01/11	1	CUA	RH	18
01/01/11	2	ACO	RH	
01/01/11	2	MON	RH	29
01/01/11	2	CHO	RH	28
01/01/11	2	CUA	RH	
01/01/11	3	ACO	RH	27
01/01/11	3	MON	RH	25.5
01/01/11	3	CHO	RH	26

```
Parser_CSV
        filas:= numero de filas en CSV
        columnas:= numero de columnas en CSV
        Generar arreglo ARG tamaño[filas]
        Inicializar i := 0
        Por cada linea en CSV
                Si linea es igual a 0 entonces
                        Verificar encabezados
                De lo contrario
                        Mandar excepción y notificar rechazo y salir de función
                Guardar la linea en [i]
                i := i+1
        Fin ciclo
        Pasar arreglo a cadena string dividiendo cada posicion con delimitar ;.
Fin Parser_CSV
```

Figura 4.3 Pseudocódigo de lectura y transformación de archivos CSV

En el caso la recolección dinámica la recuperación de información se hace por medio de un *script*, que se ejecuta cada diez minutos[18], y permite la lectura de un archivo de configuración que contiene el tiempo de actualización de cada fuente de información, así como la URL a la que se hace la petición. De tal modo a que gracias al archivo de configuración evitamos realizar peticiones a las fuentes de información cuando aún no han actualizado sus mediciones, además las respuestas que se obtienen a partir de las peticiones son válidas a partir del siguiente vector de tuplas definido como:

$$API_n = \{(k_1, V_1), \dots, (k_n, V_n)\}$$

Donde:

- **n**.- Sea el número de variables en la respuesta.
- **k**.- Sea la llave o referencia que describe la variable a recuperar.
- **V**.- Sea el valor relacionado a la llave o referencia y está definido por**:**

$$V_k = \{y | \ y \ puede \ ser \ un \ valor \ o \ un \ nuevo \ vector \ de \ tuplas\}$$

[18]Se determinó este tiempo debido a que la actualización de información se hace en diferentes tiempos como cada 20, 30 o 50 minutos.

Una vez definida nuestra respuesta válida, se procede a realizar las peticiones de manera consecutiva, obteniendo como respuesta ya sean objetos tipo JSON o XML, como se puede ver en la Figura 4.4.

{"success":true,"error":null,"response":{"id":"MMMX","loc":{"long":-99.083333333333,"lat":19.416666666667},"place":{"name":"mexico city\/lice","state":"","country":"mx"},"profile":{"tz":"America\/Mexico_City","elevM":2238,"elevFT":7343},"obTimestamp":1519919580,"obDateTime":"2018-03-01T09:53:00-06:00","ob":{"timestamp":1519919580,"dateTimeISO":"2018-03-01T09:53:00-06:00","tempC":20,"tempF":68,"dewpointC":6,"dewpointF":43,"humidity":40,"pressureMB":1009,"pressureIN":29.8,"spressureMB":782,"spressureIN":23.09,"altimeterMB":1027,"altimeterIN":30.33,"windKTS":4,"windKPH":7,"windMPH":5,"windSpeedKTS":4,"windSpeedKPH":7,"windSpeedMPH":5,"windDirDEG":100,"windDir":"E","windGustKTS":null,"windGustKPH":null,"windGustMPH":null,"flightRule":"LIFR","visibilityKM":6.437376}}

Figura 4.4 Ejemplo de respuesta de un servicio web

Una vez teniendo recolectadas la información de las diferentes fuentes tanto estáticas como dinámicas, se procede a enviar la información por medio de un *stream* de datos, que es generado utilizando la plataforma de *streaming* distribuido Apache Kafka.

En la plataforma de Apache Kafka se genera un *streaming* en el cual se puede distribuir información y mensajes, para ello se requiere que generar un tópico relacionado a nuestro dominio, por lo cual se determino generar el tópico Mediciones_Clima a las mediciones recuperadas, teniendo el tópico asociado las respuestas se envían al *stream*, el cual está siendo escuchado por el siguiente componente para su recuperación.

4.3 Pre-procesamiento de datos

En este componente se procesan los datos que se recuperan tanto de los archivos de las fuentes estáticas como de las dinámicas para su procesamiento y posterior envío al componente de integración semántica. Este proceso se realiza a través de un módulo, el cual se encargará de generar objetos que almacenen las

medidas meteorológicas y de calidad del aire con los datos recuperados como se ve en la Figura 4.5.

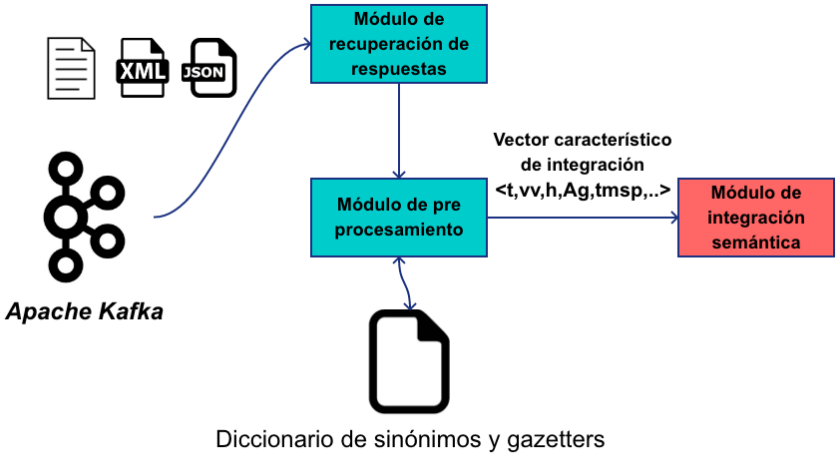


Figura 4.5 Módulo de Pre procesamiento

En el caso de las fuentes estáticas se recupera el texto plano enviado y se procede a recuperar las etiquetas asociadas a las medidas, estas se encuentran en la primera fila del texto plano, realizando una operación de separación por comas se obtiene en un arreglo las etiquetas que serán procesadas posteriormente. Al provenir la información de fuentes estáticas oficiales se consideró crear un archivo CSV histórico que funcionara como respaldo de todas las medidas recuperadas, además de servir posteriormente como conjunto de entrenamiento para el componente de análisis como se ve en la Figura 4.6.

```
Procesamiento_CSV
        Asignar  filas_proc := filas
        Asignar  col_proc := columnas
        Asignar  parse_ar := arreglo bidimensional [filas_proc][ col_proc]
        Asignar  csv := arreglo por división de texto con delimitador ;
        Asignar i :=0
        Por cada posicion en csv
                Asignar  Valores := arreglo por división de texto con delimitador ,
                Por cada posición en Valores
                        parse_ar[i][j] := Valores[j]
                        j := j+1
                Fin de ciclo
                Asignar j := 0
                Asignar i := i+1
        Fin de ciclo
Fin Procesamiento_CSV
```

Figura 4.6 Procesamiento de cadena de texto a arreglo bidimencional CSV

En el caso de las fuentes dinámicas provenientes del *stream* de datos se implementó un consumidor de *streaming,* el cual escucha la información que llega al servidor de Apache Kafka. La información recuperada llega en dos formatos XML y JSON, por lo tanto, para las respuestas de con formato JSON se crean objetos de tipo *JSONObject* para poder manipular la información; en el caso del formato XML se utiliza un *parser* que permite la búsqueda de etiquetas sobre el objeto XML.

Una vez habiendo recuperada la información dependiendo de la fuente, aún se tiene el problema de la heterogeneidad de los datos a partir de las etiquetas asignadas por el proveedor en cada fuente de datos. Un ejemplo de la heterogeneidad de nombres que presentan las diversas variables recuperadas de las fuentes de información es la medición de temperatura en grados centígrados. Esta medición es etiquetada de las siguientes maneras: temp, t, Temperature, temperature, tempC, temp_C, temp_c, CurrentTemperature.

Para abordar este problema se procede a generar un diccionario de sinónimos, conformado por los diferentes términos asignados a las mediciones realizadas por cada fuente dinámica, este diccionario se va enriqueciendo conforme se vayan añadiendo más fuentes de información.

Posteriormente, el módulo realiza las consultas al diccionario de sinónimos por las etiquetas de las variables registradas y este le devuelve un arreglo con todas las etiquetas asociadas a la variable por la que se pregunta, después utilizando una clase que modela las mediciones meteorológicas y de calidad del aire, se crean vectores que almacenan los valores recuperados de los objetos (JSON, XML, CSV), como se ve a continuación.

$$Medicion_n = < temp, vviento, hum, origen, lat, lon, fechahora, PM_{10}, O_3, SO_2, NO_2 >$$

Donde:

- **temp**.- Se refiere a la medida de temperatura, medido en grados celsius.
- **vviento**.- Se refiere a la velocidad del viento, medido en km/h.
- **hum**.- Se refiere a la cantidad de humedad en el ambiente, medido en porcentaje.
- **origen**.- Se refiere a la estación base u organización que provee las medidas.
- **lat**.- Se refiere a la latitud que posee la estación base que realizó la medición.
- **lon**.- Se refiere a la longitud que posee la estación base que realizó la medición
- **fechahora**.- Se refiere a la fecha y hora en formato timestamp en la que la medida fue realizada.
- **PM10**.- Se refiere a partículas menores a 10 micrómetros en el ambiente.
- **O3**.- Se refiere a la cantidad de ozono en el ambiente.
- **SO2**.- Se refiere a la cantidad de dióxido de azufre en el ambiente.
- **NO2**.- Se refiere a la cantidad de dióxido de nitrógeno en el ambiente.

Toda información se guarda en el vector para finalmente enviarlo al siguiente bloque y proceder a la integración semántica de las diversas fuentes de información con las que se trata en este trabajo como se ven en la Figura 4.7 y Figura 4.8.

```
Transformación_CSV
        Usar  parse_ar como base
        Asignar i :=0
        Asignar j :=1
        Asignar k :=0
        Por cada posición en el vector parser_ar[0]
                Asignar nom_var := parser_ar[0][i]
                Consultar en nom_var en diccionario de sinonimos
                Asignar pos_vec := posicion en el vector de integración desde el diccionario
                Por cada posición en el parser_ar
                        Asignar vector_integración[k][pos_vec] := parser_ar[j][pos_vec]
                        Asignar k:=k+1
                        Asignar j:=j+1
                Fin ciclo
                Asignar k:=0
                Asignar j:=1
        Fin de ciclo
        Por cada posición en vector_integración
                Mandar al módulo de integración vector_integración
        Fin ciclo
Fin Transformación _CSV
```

Figura 4.7 Transformación de arreglo CSV a vector de integración

```
Transformación_Webservices
        Si es esta en formato json entonces
                Asignar a JSONObject el valor
        De lo contrario
                Asignar a Document el valor
        Buscar en el diccionario la fuente de información
        Asignar etiqueta inicio := etiqueta que tiene los valores a recuperar
        Buscar nodo fuente y recuperar JSONObject o DocumentParent
        Por cada posición JSONObject o Document
                Asignar nombre_etiqueta := nombre de nodo
                Buscar nombre_etiqueta en el diccionario de sinonimos
                Si nombre_etiqueta tiene coincidencia entonces
                        Asignar pos_vec := posicion en el vector de integración desde el diccionario
                        Asignar vector_integración[pos_vec] := Valor de nodo
        Fin de ciclo
        Mandar al módulo de integración vector_integración
Fin Transformación_Webservices
```

Figura 4.8 Transformación de respuesta API a vector de integración

4.3 Integración semántica

Este componente se segmenta en dos partes: la primera se centra en la implementación de la red ontológica usando la metodología NeOn(Suárez et al., 2012), y la segunda parte se enfoca en la integración semántica de los datos recopilados de las diversas fuentes de información. Para ello se procede a la generación de RDF, siguiendo los principios de *Linked Data* y de la metodología (Vilches-Blázquez et al., 2014).

4.3.1 Red ontológica

Como primera parte de este componente es necesario que se tenga diseñada e implementada una red ontológica la cual nos permita realizar la integración semántica de las medidas meteorológicas y de calidad del aire. Según el trabajo de (Wache, H. Et al. 2001) se mencionan tres diferentes aproximaciones de integración basadas en ontologías:

- La primera es una aproximación basada en una única ontología global, la cual puede estar formada por varias ontologías, estas ontologías que comparten el mismo vocabulario. Este tipo de aproximación se usa mucho cuando la información a integrar es del mismo dominio.

- La segunda es una aproximación basada en múltiples ontologías donde cada una de las fuentes de información están representadas por una ontología, estas pueden estar combinadas en una sola ontología más no se puede asumir que compartan el mismo vocabulario.

- La tercera es una aproximación hibrida la cual es muy similar a la aproximación por múltiples ontologías, sin embargo, para que las ontologías locales interactúen se construye un vocabulario global el

cual contiene los términos básicos de un dominio, el vocabulario puede ser una ontología.

Debido a esto nuestra red ontológica cae en la aproximación de hibrida para la integración de fuentes de datos heterogéneas, debido a que gracias que reutilizaremos y combinaremos varias ontologías para crear la red, se construye un vocabulario global.

Para el diseño y desarrollo de la red ontológica se lleva a cabo utilizando la propuesta recogida en la metodología NeOn (Suárez, Gómez, Fernández, 2012). Las diferentes ontologías que conforman esta red únicamente poseen los elementos (conceptos, relaciones, tipos de datos y axiomas) que conforman el modelo semántico del dominio considerando; Esta estructura nos permite homogeneizar las medidas recuperadas, dándonos las relaciones de mapeo entre las medidas y los conceptos que los representan incluyendo los tipos de datos, además que las relaciones entre las ontologías nos permiten enriquecer la información integrada.

4.3.2 Generación de RDF

Tras el desarrollo de la red ontológica se procede a la materialización de la integración semántica de los diversos conjuntos de datos considerados. Para ello, se accede a la información almacenada en el vector para su integración, para realizar este proceso es necesario crear una instancia perteneciente a la red ontológica, de tal modo que el objeto sirve como entrada para el proceso, donde utilizaremos la biblioteca Apache JENA[19], dicha biblioteca nos permite crear estructuras en diferentes lenguajes como TURTLE[20], OWL[21] o RDF[22], siendo el formato RDF el que utilizaremos para la conversión, esto al juntar diferentes

[19] https://jena.apache.org/
[20] https://www.w3.org/TR/turtle/
[21] https://www.w3.org/OWL/
[22] https://www.w3.org/RDF/

formatos de datos e integrarlos acorde a los principios de *Linked Data*(Vilches-Blázquez et al., 2014).

Para continuar con el proceso como se ve en la Figura 4.9; necesitamos desplegar un *triple store*, en nuestro caso utilizaremos *Parliament*[23] como *triple store* para desplegar un *SPARQL Endpoint*, se escogió *Parliament* por el soporte que proporciona para realizar operaciones espaciales sobre las instancias. De tal modo que nuestra red ontológica estará almacenada en el *triple store*.

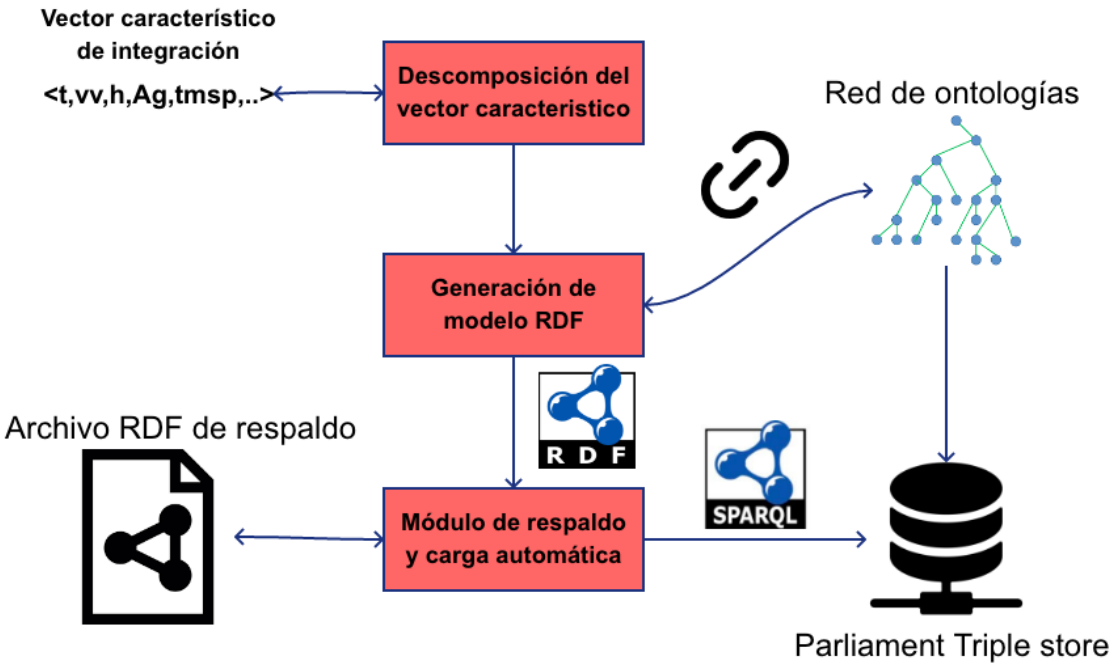


Figura 4.9. Módulo de integración semántica.

A partir de nuestra red de ontologías esta recuperaremos las URIs, conceptos, tipos de datos y relaciones con las cuales fijaremos los parámetros necesarios para crear el modelo RDF con Apache JENA. Posteriormente se crea un modelo RDF y le proporcionamos los valores del objeto y exportamos el modelo a un archivo físico en formato RDF como respaldo del resultado y al mismo tiempo realizamos una consulta de carga de la información a nuestro *SPARQL Endpoint*

[23] http://parliament.semwebcentral.org/

para que el modelo creado pase a ser integrado a nuestra red ontológica y nutra nuestro repositorio semántico de dando por finalizado el proceso de integración, esto se puede ver también en la Figura 4.10.

```
Generación_instacia
        Asignar vec_instacia := vector_integración
        Asignar Uris = Realizar consulta a la ontología por Uris
        Asignar relaciones = Realizar consulta a la ontología por relaciones
        Asignar i := número de instacia
        Asignar medicion_i := Generara nuevo esqueleto de instacia
        Por cada posición vec_instacia
                Asignar medicion_i.propiedad := vec_instacia[propiedad]
        Fin de ciclo
        Realizar consulta para actualización del SPARQL Endpoint
        Realizar archivo de respaldo de la medicón
Fin Generación_instacia
```

Figura 4.10 Generación de instancias para población de la red ontológica

4.4 Análisis

Este componente se encarga de realizar el análisis sobre el repositorio de información creado y, además, permite recuperar los resultados. Para realizar las operaciones se tiene el módulo encargado de ejecutar las operaciones sobre un conjunto de datos, para ello se realizan consultas a nuestro *SPARQL Endpoint* con el fin de recuperar las medidas que cumplan con los parámetros requeridos. El módulo realiza dos operaciones sobre nuestro *SPARQL Endpoint*, la primera es la predicción y la segunda es una comparativa entre las medidas provenientes de fuentes voluntarias y oficiales esto se ve en la Figura 4.11.

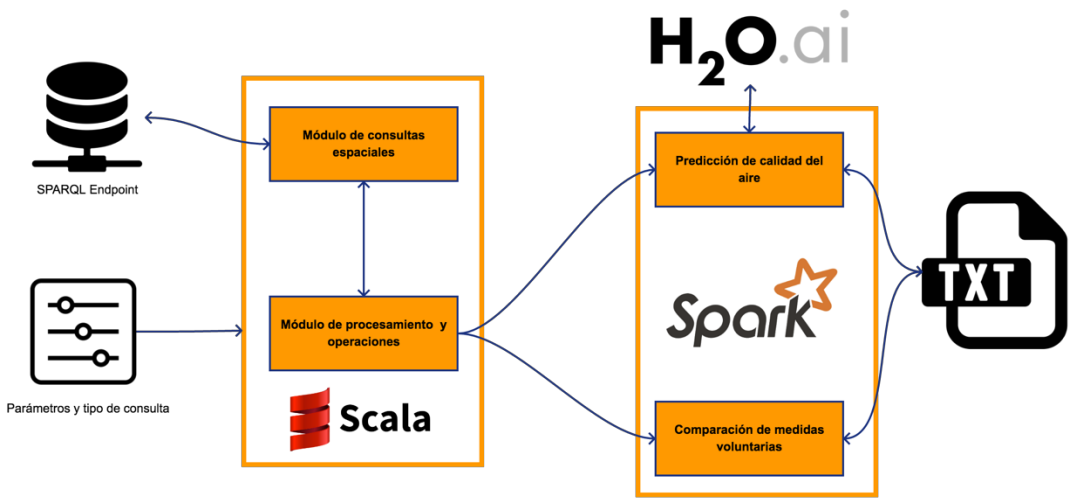


Figura 4.11. Proceso de Análisis

Para la operación de predicción, requiere de un conjunto de datos históricos con los cual se entrenará al algoritmo de predicción, de tal forma que el archivo CSV histórico de información oficial es requerido, por lo tanto, se lee el archivo el cual contiene información sobre medidas meteorológicas y de calidad del aire desde el 2011, toda la información se procesa y se transforma en RDD (*Resilient Distributed Datasets*) utilizando Apache Spark.

Una vez que se carga el repositorio se procede a construir la consulta *SPARQL* utilizando como filtro las ventanas temporales y las zonas disponibles definidas por el usuario al *SPARQL Endpoint*. La respuesta del *SPARQL Endpoint* esta en formato XML, por lo cual se procesan y se transforman en modelos RDD.

A partir de obtener los modelos RDD se hace uso de la biblioteca H2O *Sparkling Water*[24], la cual nos permite utilizar diversos algoritmos de *Machine Learning* y *Deep Learning*, que se utilizaran para analizar y realizar predicciones sobre los datos, utilizando el modelo RDD que contiene los datos históricos como conjunto de entrenamiento. Los resultados obtenidos serán mostrados al usuario en un archivo de texto donde se encontrarán las diferentes variables involucradas.

[24] https://www.h2o.ai/sparkling-water/

Para la operación de comparación de la calidad de los datos se realizan dos consultas a nuestro repositorio de datos, la primera para recuperar del grafo donde están almacenados todos los datos de fuentes voluntarias y una segunda al grafo donde se encuentra la información oficial. Ambas consultas se realizan por ventanas temporales que son definidas por el usuario, proveedores de información y por las zonas disponibles en el repositorio de datos. Teniendo ambas respuestas, se transforman a modelos RDD para realizar las operaciones sobre ellos y determinar si hay discrepancias entre las medidas provenientes de fuentes las diversas fuentes (oficiales y voluntarias).

CAPÍTULO 5. EXPERIMENTOS Y RESULTADOS

En este capítulo se describen las pruebas de la pertinencia y validez de la metodología descrita en el Capítulo 4. Se usa un caso de estudio con la finalidad de validar las funcionalidades de la metodología.

La sección se compone de la descripción detallada del caso de estudio, el diseño de la red ontológica empleada los escenarios de prueba utilizados y finalmente la integración semántica.

5.1 Caso de Estudio

El caso de estudio considerado para este trabajo consiste en el estudio de las estaciones base que se encuentran dentro de la Ciudad de México (CDMX), asociadas a fuentes oficiales y voluntarias. La Ciudad de México cuenta con un Sistema de Monitoreo Atmosférico conformado por una red de monitoreo vigente desde el año 1986 hasta la actualidad. Dicha red fue reformada en el año 2011 para incluir la calidad del aire entre sus mediciones. Actualmente la red cuenta con 28 estaciones meteorológicas y de calidad del aire.

En la figura 5.1 se muestra la división de los sectores del Sistema de Monitoreo Atmosférico de la Ciudad de México, conformado por los siguientes sectores: Noroeste (NO), Noreste (NE), Sureste (SE), Suroeste (SO) y Centro. Las estaciones meteorológicas se dividen en estas secciones y recolectan las mediciones cada hora.

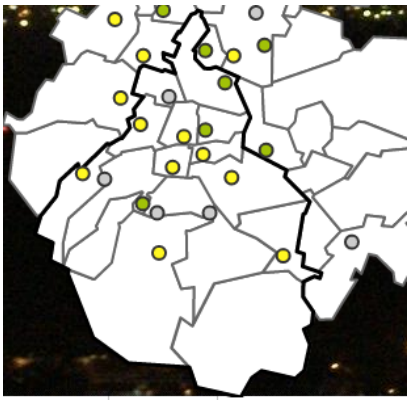


Figura 5.1. Mapa de calidad del aire del Sistema de Monitoreo Atmosférico de la Ciudad de México[25].

Dependiendo el proveedor son las estaciones asociadas a él, por ejemplo, las fuentes voluntarias como *WeatherUnderground* poseen 14 estaciones en la CDMX como se pueden ver en la figura 5.2.

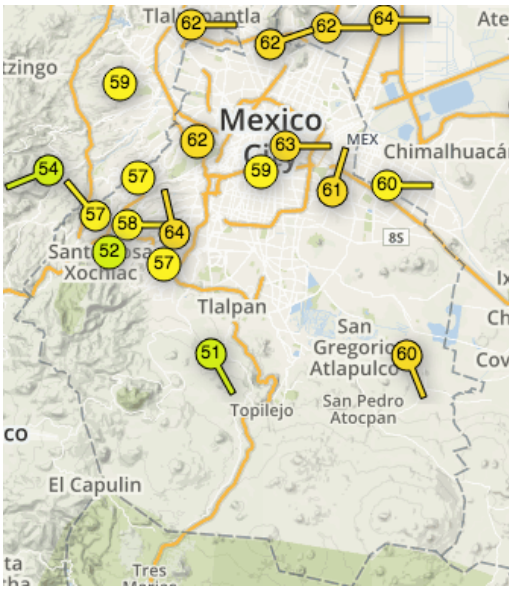


Figura 5.2. Mapa de ubicación de las estaciones base en la CDMX registradas en *WeatherUnderground*.

En las siguientes subsecciones se describirá la implementación de la metodología, además, el caso de estudio expuesto con la finalidad de demostrar la validez de la misma.

[25] http://www.aire.cdmx.gob.mx/default.php

5.2. Fuentes de información

Las fuentes de información tanto oficiales como voluntarias poseen diversas características, como son: el tipo de respuesta, el formato en el que responden, las etiquetas que manejan para cada variable, la cobertura territorial que poseen y si tienen un servicio web asociado para consulta. Estas características las podemos ver en la tabla 5.1

Tabla 5.1. Características de las fuentes de información consultadas

Fuente de información	Tipo de fuente	Formato de resp.	Web Service	Covertura	Unidades	Limitaciones
CONAGUA	Oficial (MX)	CSV	NO	México		
REDMET	Oficial (MX)	CSV	NO	CDMX	RH, TMP, WDR, WSP	La información la proporcionan anual
SEMAR	Oficial (MX)	CSV	NO	Costas de mexico y CDMX	Temp, Hr, PCP Dirmx	Se requiere consultar estación por estación
3TIER	Voluntaria	JSON	RESTful	NI	NI	Es de paga
AccuWeather	Voluntaria	JSON	RESTful	Mundial	Temperature, Imperial	Cantidad Limitada de peticiones
FAA Airport Service	Oficial (EUA)	XML y JSON	RESTful	Solo EUA	Visibility, temp, wind	Solo los aeropuertos de EUA por su ITA Code
Foreca	Voluntaria	XML y JSON	RESTful	Mundial	diccionario de simbolos	Solo dan 100 queries gratis después se pegan 2000€

GEONAMES	Voluntaria	Varios Lenguajes	RESTFul	Mundial	temperature, humidity, windDirection, windSpeed	Ninguna
Aerisapi	Voluntaria	XML y JSON	RESTful	Mundial	tempC, tempF, humidity, , windMPH, windSpeedKPHwin dDir,	Hay que registrarse para tener acceso limitado.
Weatherbug api pulse	Voluntaria	JSON	RESTful	Mundial	NI	La estan renovando
WeatherUndergr ound	Voluntaria	JSON	RESTful	Mundial	temp_f, temp_c, relative_humidity, wind_string, wind_dir, wind_mph	Solo 30 por mes, se necesita llave apartir de registro.

5.3. Diseño de red ontológica

Para la creación de la red de ontologías se utiliza la metodología *NeOn*, la cual plantea nueve diferentes escenarios que son aplicados conforme a las necesidades que se requieren cubrir con una red de ontologías. Por este motivo se proponen los escenarios 3, 4 y 7 para el diseño de la red ontológica.

De acuerdo a lo planteado en el trabajo de (Poveda, 2009), se utilizará la plantilla para el documento de especificación de requisitos de la ontología *ORSD (Ontology Requirements Specification Document)* descrita en (Suárez-Figueroa et al., 2008), con el fin de ayudar a identificar los dominios ontológicos necesarios, la plantilla se conforma de las siguientes subsecciones:

Propósito: El propósito de la red de ontologías es realizar la integración de diferentes fuentes heterogéneas de información meteorológica y de calidad del aire, ya sean de carácter voluntario u oficial, dinámicas o estáticas.

Alcance: La red de ontologías usa información relacionada a los siguientes dominios:

- Condiciones meteorológicas y atmosféricas: Se enfoca en las variables involucradas con el tiempo de una región, desde la temperatura hasta el rango de humedad de un lugar, así como todos los contaminantes que están presentes en el aire y que, en determinadas condiciones o concentraciones, son dañinas para el ser humano.

- Geografía: Se refiere a la información de la localización geográfica de objetos en un área o áreas afectadas por fenómenos meteorológicos o por contaminantes atmosféricos.

- Sensores: Trata acerca de las mediciones obtenidas por medio de un sensor, los tipos de sensores y el tiempo de medición entre otras.

- Procedencia: Este domino se refiere al origen de los datos, así como de las organizaciones que producen la información.

Nivel de formalidad: Para este trabajo, es necesario que la ontología esté en un lenguaje *OWL-Full* debido a que forma parte de las recomendaciones de la *W3C*.

Usuarios previstos: Los usuarios a los que va dirigida la red son investigadores que estén interesados en datos meteorológicos integrados o en realizar análisis sobre ellos.

Usos previstos: Permitir la integración de diferentes fuentes de información, así como la consulta y análisis sobre los datos integrados.

Grupos de competencia: Las preguntas de competencia que la red ontológica responderá son:

- Condiciones Meteorológicas y Atmosféricas

    - 1.- ¿Cómo está el tiempo actualmente?: Se proporcionarán las medidas actuales del tiempo.
    - 2.- ¿Está muy contaminado el aire?: Se proporcionarán las cantidades de contaminantes en el aire que determinen si la calidad es buena, regular, mala, muy mala o extremadamente mala.
    - 3.- ¿Cuál será la calidad del aire para mañana?: Se proporcionarán las cantidades predichas de los contaminantes en el aire, con base en los datos históricos recolectados.

- Geográfica

    - 4.- ¿Dónde están las estaciones base que usan?: Se proporcionarán las localizaciones de las estaciones base que posean posición geográfica.
    - 5.- ¿Dónde se localizan las estaciones que pertenecen a fuentes oficiales o voluntarias?: Se proporcionarán las localizaciones de las estaciones base relacionadas a su tipo de fuente.
    - 6.- ¿Cómo está el tiempo en mi zona?: Se proporcionarán las medidas meteorológicas con base en la zona solicitada.
    - 7.- ¿A qué zona pertenecen las medidas recuperadas?: Se proporcionarán las medidas relacionadas a la zona de búsqueda.

- Sensores

    - 8.- ¿Cuál es la temperatura?: Se proporcionará la medición de la temperatura requerida por el usuario.
    - 9.- ¿Cuál es la cantidad de ozono hoy a las 10 am en la zona noroeste?: Se proporcionará la cantidad de ozono requerida con base a estos parámetros.
    - 10.- ¿Cuándo se tomó la medición?: Se proporcionará el *timestamp* asociado a la medición.

- Proveniencia

    - 11- ¿Quién proporcionó esta información?: Se proporcionará el nombre del productor de los datos.
    - 12.- ¿Qué tipo de fuente proporcionó la información?: La fuente puede ser voluntaria u oficial.

5.3.1 Escenario 3: La reutilización de recursos ontológicos.

Debido a que reutilizamos los recursos de las ontologías, sin hacer reingeniería ontológica, se utilizará el escenario 3, por lo cual se procede a describir las ontologías que son reutilizadas en el proceso de construcción de esta red de ontologías.

- Prov-O[26]: Esta especificación de ontología provee los principios para implementar la procedencia de aplicaciones en diferentes dominios que puedan representar, intercambiar e integrar la procedencia de la información generada en diferentes sistemas y bajo diferentes contextos. Para nuestra red de ontologías, la *Prov-O* permite almacenar la información del tiempo en

[26] https://www.w3.org/TR/prov-o/

el que se realizó la medición, así como el Agente que la está produciendo. En el caso de los servicios *Web* y *APIs,* son los proveedores del servicio, y, en el caso de las estaciones base voluntarias, son los dueños que permiten el libre uso de sus mediciones.

- *GeoSPARQL*[27]: Es un estándar de la *OGC*, el cual define una extensión de funciones para *SPARQL* (W3r), un conjunto de reglas *RIF*[28] y un núcleo *RDF/OWL*[29] el vocabulario para la información geográfica basada en el *General Feature Model* (ISO 19109), *Simple Features* (ISO 19125), *Feature Geometry* (ISO 19107) y SQL MM (ISO/IEC 13249-3). Para nuestra red ontológica, esta ontología permite el uso de componentes espaciales (coordenadas geográficas de las estaciones base), así como la posibilidad de realizar operaciones espaciales sobre las instancias almacenadas en el *EndPoint*, utilizando las coordenadas en *GML* o en formato *WKT*.

5.3.2 Escenario 4: La reutilización y re-ingeniería de los recursos ontológicos

Debido a que reutilizamos los recursos de las ontologías, haciendo reingeniería sobre ellas, se utilizará el escenario 4. A continuación, se describen las ontologías que se modifican:

- *Semantic Sensor Network(SSN)*[30]*:* Ésta es una ontología para describir sensores y sus observaciones, los procesos involucrados, el estudio de las características de interés, los ejemplos empleados y la observación de propiedades al igual que sus actuadores. Para nuestra red de ontologías, esta ontología le da la capacidad a la red de registrar las propiedades que se van a observar, como son: la

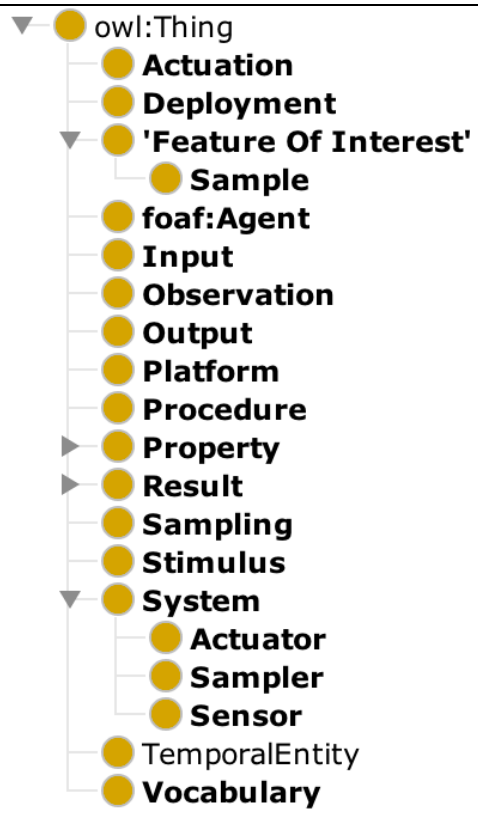
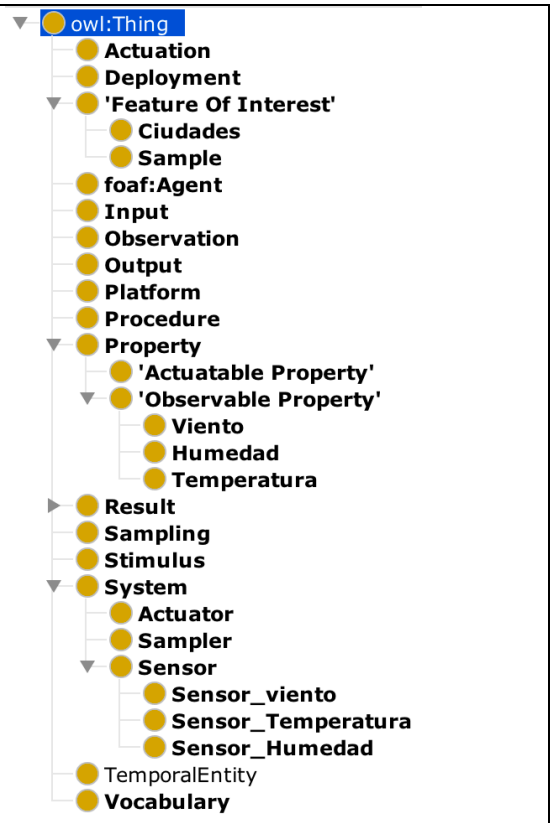
[27] http://www.opengeospatial.org/standards/geosparql
[28] http://www.w3.org/TR/rif-core/
[29] http://www.w3.org/RDF/
[30] https://www.w3.org/TR/vocab-ssn/

humedad, la temperatura y la velocidad del viento. Al igual que los valores de las mismas y los tipos de datos. Conforme al *XMLSchema* y los conceptos que se van a observar, de manera similar a la ontología *Prov-O*, se almacena el tiempo en el que se realizó la medición.

- *Sensor*, *Observation*, *Sampler and Actuator(SOSA)*[31]: Esta ontología actúa como bloque constructor para la SSN pero enfatiza en su bajo peso y la habilidad ser usada en entornos *standalone*.

Nombre	Completar conocimiento representado
Nivel	Re-conceptualización
Descripción	
Esta modificación consiste en ampliar el conocimiento representado añadiendo elementos como clases a las ontologías SOSA y SSN. En particular, se añaden clases a los conceptos de '*Feature of interest*', '*Sensor*', '*Observable Property*'	
Ejemplo	
En la situación de los '*Feature of interest*' se requiere definir clases sobre las cuales se realizan las mediciones, debido a los diferentes niveles de granularidad que se tienen.	
Situación Inicial	Situación Final

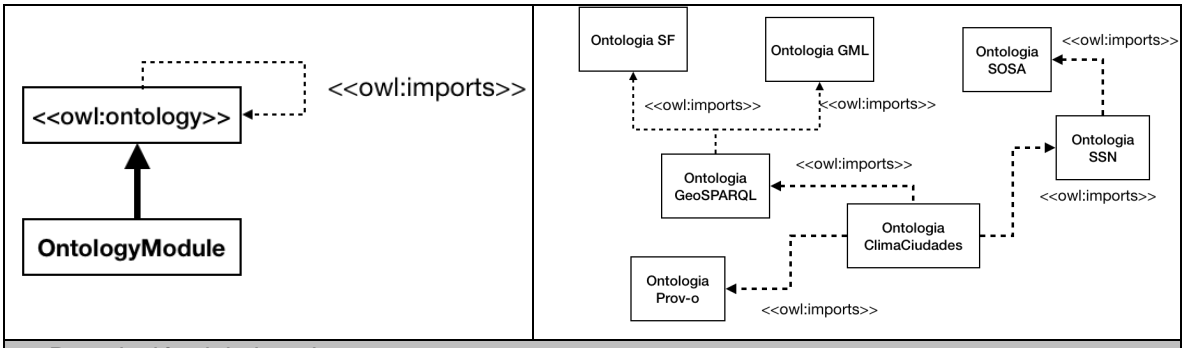
[31] https://www.w3.org/TR/vocab-ssn/

	
Motivación	
La motivación principal para realizar esta modificación es añadir conocimiento no representado que forme parte del dominio y de los requisitos de la ontología.	

5.3.3 Escenario 7: Reutilización de los patrones de diseño de ontologías (ODPs)

Durante el desarrollo de la ontología se han reutilizado algunos de los patrones de diseño, particularmente la importación directa de las diferentes ontologías, así como importaciones indirectas de cada ontología. Por esta razón, la utilización del escenario 7 es necesario, como se presenta a continuación:

Nombre	Arquitectura Modular		
Origen	Propio	Identificador	AP-MD-01
Diagrama para la solución general		Ejemplo grafico	



Descripción del ejemplo

En este ejemplo se puede observar las ontologías importadas directamente. Las cuales son: *Prov-o, GeoSPARQL y Semantic Sensor Network (SSN)*, e indirectamente como son las ontologías *Simple Features, Geography Markup Language(GML)* y *Sensor,Observation, Sampler and Actuator (SOSA)* en la nueva ontología Clima Ciudades.

Una vez concluida la aplicación de los escenarios, se tiene como resultado la red ontológica, la cual es la base para llevar a cabo el proceso de integración como se muestra en la Figura 5.3.

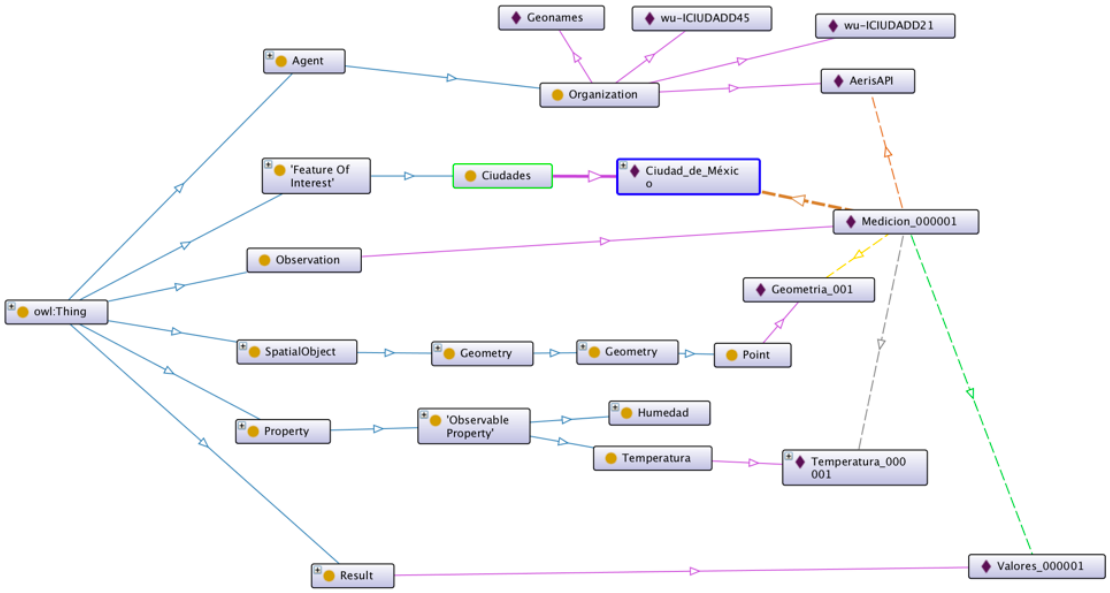


Figura 5.3. Vista en *OntoGraph* de la red ontológica

5.4 Proceso de integración de fuentes heterogéneas

Como se describe en el capítulo 4, a continuación se ejemplifica el proceso de integración desde la recolección de datos hasta el análisis.

5.4.1 Recolección de datos

El proceso inicia a través de la ejecución del script de inicialización del *zookeeper*[32] el cual es un servicio centralizado para mantener la información configurada y nombrada, la cual proporciona la sincronización distribuida y servicios grupales. Posteriormente, se inicializa el servicio de *Apache Kafka* que conecta con el script *zookeeper*, el servicio utiliza el puerto 9092 para generar la distribución y recepción de datos. Los comandos utilizados para estas tareas son:

- bin/zookeeper-server-start.sh config/zookeeper.properties
- bin/kafka-server-start.sh config/server.properties

Una vez inicializado el servicio, se procede a ejecutar el script encargado de la recolección de datos. Este servicio está desarrollado en lenguaje JAVA, y está dividido en dos micro servicios, los cuales son:

- Recuperación de fuentes dinámicas: en este servicio se utiliza un archivo de configuración que posee las peticiones que se realizarán por cada estación base, además del tiempo de actualización de cada proveedor. A continuación, se muestra un ejemplo de fuentes disponibles en sistema en la Figura 5.4.

[32] https://zookeeper.apache.org/

```
aerisapi;10;http://api.aerisapi.com/observations/guadalajara,mx?client_id=SqkoJJCJOTGpipZWoXfGJ&client_secret=GWlgvg5TlTgkNgdCXiztupsm8NG0unEo1HDQO8PR
geonames;50;http://api.geonames.org/findNearByWeatherJSON?lat=20.659698&lng=-103.349609&username=fausto29
wu-ICIUDADD21;20;http://api.wunderground.com/api/af35041da16e2b99/conditions/q/pws:ICIUDADD21.json
wu-ICIUDADD45;http://api.wunderground.com/api/af35041da16e2b99/conditions/q/pws:ICIUDADD45.json
```

Figura 5.4. Ejemplo del archivo de configuración

Posteriormente, el *script* ejecuta una rutina temporizada, la cual se ejecuta cada 10 minutos y está encarga de realizar peticiones a los servicios web que aparecen en el archivo de configuración dependiendo de su tiempo de actualización. Las peticiones son de tipo GET, debido a que no son enviados parámetros a los servicios web y su respuesta varía de formato dependiendo del proveedor, por lo cual se convierte a texto plano y se envían al *stream* de datos utilizando el método mostrado en la Figura 5.5, dando por terminado el primer servicio.

```
private static final String KAFKA_SERVER = "localhost:9092";

public ConfigKafka() {
        final Properties props = new Properties();
        props.put("metadata.broker.list", KAFKA_SERVER);
        props.put("client.id", "testClient");
        props.put("serializer.class", "kafka.serializer.StringEncoder");
        producer = new Producer<String, String>(new ProducerConfig(props));
    }

    public void send(String topic, String message) {
        producer.send(new KeyedMessage<String, String>(topic, message));
    }
```

Figura 5.5. Conexión y envío de datos a Apache Kafka

- El segundo servicio se enfoca en la recuperación de las fuentes de datos estáticas que utilizan formato CSV. En la Tabla 5.2 se muestra un fragmento de un archivo CSV. Para realizar esta recuperación de datos se utiliza una rutina encargada de leer y procesar la información del archivo, teniendo como parámetro de entrada, la línea donde se encuentran las etiquetas de las mediciones. El archivo se lee a partir de la línea indicada hacia abajo, para

almacenar toda la información en un texto plano que se enviará al *stream* utilizando la misma función presentada en la Figura 5.5.

Tabla 5.2. Fragmento de un archivo CSV

INDICE DE CALIDAD DEL AIRE						
city: Ciudad de México						
cityCode: MEX						
country: México						
mesurementAgency: SIMAT						
URL: http://www.aire.cdmx.gob.mx						
timeStamp: 2017/01/01 al 2017/12/31						
average_interval: 001h						
Fecha	Hora	Noroeste Ozono	Noroeste dióxido de azufre	Noroeste dióxido de nitrógeno	Noroeste monóxido de carbono	Noroeste PM10
01/01/17	1	9	18	18	12	80
01/01/17	2	5	15	17	13	86
01/01/17	3	6	12	16	13	94
01/01/17	4	3	9	17	13	100

5.4.2 Pre Procesamiento

El componente del pre-procesamiento se describirá a continuación conforme a lo descrito en el capítulo 4.

5.4.2.1 Módulo de recuperación de información

El módulo de recuperación se implementa en lenguaje JAVA e inicia ejecutando el *script* encargado de estar escuchando el *stream* de datos. Cuando se recibe una actualización en el tópico Mediciones_Clima (definido en el capitulo 4), el *script* lo recupera y lo almacena en una variable de tipo *String* como se ve en la Figura 5.6.

a)

{"success":true,"error":null,"response":{"id":"MMMX","loc":{"long":-99.083333333333,"lat":19.416666666667},"place":{"name":"mexico city\/lice","state":"","country":"mx"},"profile":{"tz":"America\/Mexico_City","elevM":2238,"elevFT":7343},"obTimestamp":1519919580,"obDateTime":"2018-03-01T09:53:00-06:00","ob":{"timestamp":1519919580,"dateTimeISO":"2018-03-01T09:53:00-06:00","tempC":20,"tempF":68,"dewpointC":6,"dewpointF":43,"humidity":40,"pressureMB":1009,"pressureIN":29.8,"spressureMB":782,"spressureIN":23.09,"altimeterMB":1027,"altimeterIN":30.33,"windKTS":4,"windKPH":7,"windMPH":5,"windSpeedKTS":4,"windSpeedKPH":7,"windSpeedMPH":5,"windDirDEG":100,"windDir":"E","windGustKTS":null,"windGustKPH":null,"windGustMPH":null,"flightRule":"LIFR","visibilityKM":6.437376}}

b)

Fecha,Hora,Noroeste Ozono,Noroeste dióxido de azufre,Noroeste dióxido de nitrógeno,Noroeste monóxido de carbono,Noroeste PM10,
01/01/2017,1,9,18,18,12,
01/01/2017,2,5,15,17,13,
01/01/2017,3,6,12,16,13,
01/01/2017,4,3,9,17,13,
01/01/2017,5,3,8,16,13,

Figura 5.6. Información recibidos del *stream* de datos a) fuente dinámica, b) fuente estática.

El *script* procede a clasificar el tipo de formato que posee la información que contiene la variable tipo *String*, para poder crear los objetos con los se extraerá la información teniendo los formatos dinámicos (JSON, XML) y los estáticos (CSV). Comenzando por el caso de las fuentes estáticas, donde la variable de tipo *String* contiene texto plano separado por comas, esta información se almacena en un arreglo bidimensional y es separada utilizando como delimitador las comas, después es enviado al módulo de procesamiento.

El segundo caso son las fuentes dinámicas donde se tienen registrados dos formatos de respuesta (JSON, XML), por lo tanto, la variable de tipo *String* se intenta convertir a un *JSONObject* que contenga la información. Si la operación presenta una excepción de formato, automáticamente la variable es enviada a un *parser* de XML para obtener un objeto de tipo *Document* que almacene la información. Una vez obtenido el objeto que almacena la información se manda al módulo de procesamiento.

5.4.2.2 Módulo de procesamiento de información

El módulo recibe los objetos (*JSONObject*, *Document* o Arreglo) para procesarlos y extraer las mediciones meteorológicas y de calidad del aire que contengan, para ello en el capítulo 4 se describe el diccionario de sinónimos, así como su función. El proceso inicia cuando llega un objeto y, dependiendo de su formato, se determina que método lo procesará. Sin embargo, independiente del formato, todos los métodos realizan primero la consulta al diccionario de datos dependiendo de la variable que se quiera extraer en ese momento, tal como se puede observar en la Figura 5.7, en donde se recuperan los términos relacionados con la temperatura.

```
Object obj = parser.parse(new
FileReader("/Users/luich/Documents/Tesis/data2RDF/variables.json"));

        //Variables de respuesta
        JSONObject jsonObject = (JSONObject) obj;
        JSONObject recursos = (JSONObject) jsonObject.get("inicio");
        //System.out.println(recursos.toString());

        JSONObject variables = (JSONObject) jsonObject.get("variables");
        JSONArray temperatura = (JSONArray) variables.get("temperatura");
```

Figura 5.7. Recuperación de sinónimos de temperatura.

Teniendo el arreglo de sinónimos de la variable a consultar, se realiza la búsqueda basada en los términos asociados a la variable en cuestión sobre las etiquetas del objeto de la etiqueta, en este caso a la temperatura. Después de recuperar todas las variables de los objetos, se genera un vector estándar a partir de las variables recuperadas como se ve en le Figura 5.8.

```
<”32”,”06”,”25”,”geonames”,” 20.516666666666666”,”103.31666666666666”,” 2018-06-
07T17:46:00-05:00”,”0”,”0”,”0”,”0”>
```

Figura 5.8. Vector creado a partir de las variables recuperadas.

Para finalizar el proceso, se considera el tipo de fuente de información, si es una fuente dinámica, únicamente se manda al siguiente componente el vector que representa; si la fuente fue dinámica, se procede a mandar el vector que represente y además se realiza una escritura sobre el archivo histórico de medidas. Esto, al

descartar la primera fila que contiene los nombres de las etiquetas y asignar a los espacios pertenecientes los valores recuperados. Las variables que están presentes en la información recuperada se completan con espacios vacíos, como se ve en la Figura 5.9.

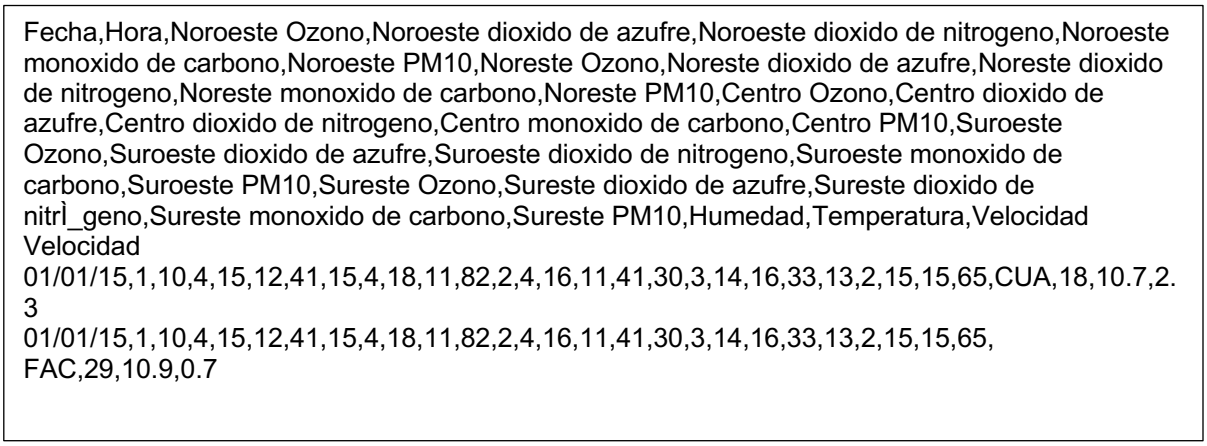


Figura 5.9. Ejemplo de actualización del archivo histórico.

5.4.3 Integración semántica

Este módulo requiere la red de ontologías para integrar los datos guardados en los objetos procesados, con la finalidad de generar un modelo RDF que será integrado al repositorio de datos que se encuentra en nuestro *SPARQL Endpoint*, para ello se utiliza el objeto generado en el componente anterior. Para esto se construye una clase que modela una instancia perteneciente a la red de ontologías, como se puede ver en la Figura 5.10.

```
private String proveniencia;
  private String temperatura;
  private String latitud;
  private String longitud;
  private String viento;
  private String humedad;
  private String offset;
  private String fechaToma;

  public Instancia(String proveniencia, String temperatura, String latitud, String longitud, String
viento, String humedad, String offset, String fechaToma) {
    this.proveniencia = proveniencia;
    this.temperatura = temperatura;
    this.latitud = latitud;
    this.longitud = longitud;
    this.viento = viento;
    this.humedad = humedad;
    this.offset = offset;
    this.fechaToma = fechaToma;
  }
}
```

Figura 5.10. Clase basada en las propiedades de una instancia de la red ontológica.

Una vez teniendo el modelo, se procede a utilizar la biblioteca JENA para construir el modelo RDF, que es enviado a poblar la red. Además de generar un archivo de respaldo, para esto se recuperan las URIs que utiliza la red de ontologías, teniendo en cuenta que los dominios que se usan son:

- http://datos.climaticosCiudades.mx/ontologia/
  para los conceptos agregados a la red ontológica en el escenario 4.

- http://datos.climaticosCiudades.mx/recursos/geometry/
  para la geometría que puedan tener las mediciones.

- http://www.opengis.net/ont/sf#Point
  para definir las mediciones como geometrías puntuales sobre un mapa.

- http://www.opengis.net/ont/geosparql#
  para definir los recursos de GeoSPARQL.

- http://www.w3.org/ns/prov#
  para definir de donde provienen los datos, así como la organización que se encarga de producirlos.

- http://www.w3.org/ns/sosa/
  para definir las mediciones y las propiedades que se recuperan.

- http://www.w3.org/ns/ssn/

define los procedimientos, las propiedades observables, las características de interés y los resultados de los procedimientos.

Una vez recuperadas las URIs, se extraen las variables contenidas en el vector y se asignan a las propiedades del modelo RDF, como se ve en la Figura 5.11.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:prov="http://www.w3.org/ns/prov#"
  xmlns:sosa="http://www.w3.org/ns/sosa/"
  xmlns="http://datos.climaticosCiudades.mx/ontologia/"
  xmlns:geo="http://datos.climaticosCiudades.mx/recursos/geometry/"
  xmlns:geosparql="http://www.opengis.net/ont/geosparql#"
  xmlns:ssn="http://www.w3.org/ns/ssn/"
  xmlns:j.0="http://www.opengis.net/ont/sf#">
 <sosa:Observation rdf:about="http://datos.climaticosCiudades.mx/recurso/Medicion_3">
  <prov:wasAttributedTo rdf:resource="http://datos.climaticosCiudades.mx/ontologia/geonames"/>
  <geosparql:hasGeometry>
   <j.0:Point rdf:about="http://datos.climaticosCiudades.mx/recurso/Geometria_3">
    <geosparql:asGML rdf:datatype="http://www.opengis.net/ont/geosparql#gmlLiteral"
    ><![CDATA[<gml:Point srsName="EPSG:4326" xmlns:gml="http://www.opengis.net/gml"><gml:coordinates decimal="." cs="," ts=" ">-103.31666666666666,20.516666666666666 </gml:coordinates></gml:Point>]]></geosparql:asGML>
    <geosparql:asWKT rdf:datatype="http://www.opengis.net/ont/sf#wktLiteral"
    ><![CDATA[<http://www.opengis.net/def/crs/OGC/1.3/CRS84>Point(-103.31666666666666
20.516666666666666)]]></geosparql:asWKT>
   </j.0:Point>
  </geosparql:hasGeometry>
  <sosa:observedProperty>
   <Temperatura rdf:about="http://datos.climaticosCiudades.mx/ontologia/Temperatura_3">
    <ssn:isProperty rdf:resource="http://datos.climaticosCiudades.mx/ontologia/Ciudad_de_México"/>
   </Temperatura>
  </sosa:observedProperty>
  <sosa:hasResult>
   <sosa:Result rdf:about="http://datos.climaticosCiudades.mx/ontologia/Valores_3">
    <sosa:resultTime rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime"
    >2018-06-07T17:46:00-05:00</sosa:resultTime>
    <prov:atTime rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime"
    >2018-06-07T17:46:00-05:00</prov:atTime>
    <MedicionTemperatura rdf:datatype="http://www.w3.org/2001/XMLSchema#double"
    >32.0</MedicionTemperatura>
    <MedicionHumedad rdf:datatype="http://www.w3.org/2001/XMLSchema#double"
    >25.0</MedicionHumedad>
    <MedicionViento rdf:datatype="http://www.w3.org/2001/XMLSchema#double"
    >6.0</MedicionViento>
   </sosa:Result>
  </sosa:hasResult>
  <sosa:hasFeatureOfInterest rdf:resource="http://datos.climaticosCiudades.mx/ontologia/Ciudad_de_México"/>
 </sosa:Observation>
 <Humedad rdf:about="http://datos.climaticosCiudades.mx/ontologia/Humedad_3">
  <ssn:isProperty rdf:resource="http://datos.climaticosCiudades.mx/ontologia/Ciudad_de_México"/>
 </Humedad>
 <Viento rdf:about="http://datos.climaticosCiudades.mx/ontologia/Viento_3">
  <ssn:isProperty rdf:resource="http://datos.climaticosCiudades.mx/ontologia/Ciudad_de_México"/>
 </Viento>
</rdf:RDF>
```

Figura 5.11 Salida del modelo RDF.

Finalmente, al tener el modelo RDF se genera el archivo de respaldo de la medida y al mismo tiempo la consulta de carga a *SPARQL Endpoint* Figura 5.12.

```
String queryString = "";
      queryString = "PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>"
            + "PREFIX fn: <http://www.w3.org/2005/xpath-functions#>"
            + "PREFIX geo: <http://www.opengis.net/ont/geosparql#>"
            + "PREFIX geof: <http://www.opengis.net/def/function/geosparql/>"
            + "PREFIX gml: <http://www.opengis.net/ont/gml#>"
            + "PREFIX owl: <http://www.w3.org/2002/07/owl#>"
            + "PREFIX par: <http://parliament.semwebcentral.org/parliament#>"
            + "PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>"
            + "PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>"
            + "PREFIX sf: <http://www.opengis.net/ont/sf#>"
            + "PREFIX time: <http://www.w3.org/2006/time#>"
            + "PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>"
            + "PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>"
+ "LOAD <file:///Users/luich/Documents/Tesis/instanciaWeatherMeasure_5taPasada_"+inst.getOffset()+".rdf>";

      UpdateRequest update = UpdateFactory.create(queryString);
      UpdateProcessor processor =
                  UpdateExecutionFactory.createRemote(update, "http://localhost:8089/parliament/sparql");
      processor.execute();
```

Figura 5.12. Consulta de carga de archivos a un *SPARQL Endpoint*.

5.4.4 Análisis

Para el análisis se crea un programa en el lenguaje SCALA[33], que hace uso de las bibliotecas: *Apache SPARK[34], Apache JENA* y *H2O Sparkling water*, como base para los procedimientos a realizar. El programa recibe del usuario los parámetros que sirven para filtrar las consultas realizadas al *SPARQL Endpoint*, teniendo la capacidad de realizar operaciones espaciales utilizando la ontología *GeoSPARQL* que se encuentra integrada en nuestra red de ontologías, las consultas pueden ser de la siguiente manera:

- Por rango de tiempo.- Se recupera la fecha y hora de inicio, además de la fecha y hora de fin del rango temporal, esta fecha se proporciona usando el tipo de dato dateTime de XMLSchema, como *default* se pregunta por la mediciones pertenecientes a la instancia de Ciudad_de_México, a partir de ellos se preguntan por los valores asociados a las instancias, las variables meteorológicas, de calidad del aire y la ubicación, la cual está

[33] https://www.scala-lang.org/
[34] https://spark.apache.org/

definida por la existencia de la relación hasGeometry al concepto de geometría Figura 5.13.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
    PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
    PREFIX geo: <http://www.opengis.net/ont/geosparql#>
    PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
    PREFIX gml: <http://www.opengis.net/ont/gml#>
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    PREFIX par: <http://parliament.semwebcentral.org/parliament#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    PREFIX sf: <http://www.opengis.net/ont/sf#>
    PREFIX time: <http://www.w3.org/2006/time#>
    PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
    PREFIX xml: <http://www.w3.org/XML/1998/namespace>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX prov: <http://www.w3.org/ns/prov#>
    PREFIX sosa: <http://www.w3.org/ns/sosa/>
    PREFIX onto: <http://datos.climaticosCiudades.mx/ontologia/>
    PREFIX my:<http://datos.itinerariosTuristicos.mx/recursos/>
    PREFIX geome:<http://datos.climaticosCiudades.mx/recursos/geometry/>

     SELECT DISTINCT *
     WHERE {
        ?mediciones sosa:hasFeatureOfInterest onto:Ciudad_de_México.
        ?mediciones sosa:hasResult ?resultados.
        ?mediciones prov:wasAttributedTo ?Agente.
        ?resultados prov:atTime ?tiempo.
         FILTER(?tiempo > "2018-04-19T15:00:00-05:00"^^xsd:dateTime && ?tiempo < "2018-04-19T23:00:00-
05:00"^^xsd:dateTime).
       ?resultados onto:MedicionTemperatura ?temperatura
    }
```

Figura 5.13 Consulta para recuperar información por rango de tiempo únicamente.

- Por localización y tiempo.- Este filtro puede utilizarse con dos viables: filtrado por una de las zonas en las que esta segmentado la Ciudad de México e, indicando una ubicación, la cual servirá como centro para realizar operaciones espaciales, en este caso se define un *buffer* de 1 km alrededor de la ubicación indicada y, de esta manera, se recuperan las estaciones bases contenidas en el *buffer*. En ambos casos se aplica la ventana de tiempo para recuperar solo las mediciones de la hora y/o fecha. En la Figura 5.14 se muestra la consulta para recuperar información por rango de tiempo y ubicación.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
    PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
    PREFIX geo: <http://www.opengis.net/ont/geosparql#>
    PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
    PREFIX gml: <http://www.opengis.net/ont/gml#>
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    PREFIX par: <http://parliament.semwebcentral.org/parliament#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    PREFIX sf: <http://www.opengis.net/ont/sf#>
    PREFIX time: <http://www.w3.org/2006/time#>
    PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
    PREFIX xml: <http://www.w3.org/XML/1998/namespace>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX prov: <http://www.w3.org/ns/prov#>
    PREFIX sosa: <http://www.w3.org/ns/sosa/>
    PREFIX onto: <http://datos.climaticosCiudades.mx/ontologia/>
    PREFIX my:<http://datos.itinerariosTuristicos.mx/recursos/>
    PREFIX geome:<http://datos.climaticosCiudades.mx/recursos/geometry/>
    PREFIX geosparql:<http://www.opengis.net/ont/geosparql#>

     SELECT DISTINCT *
     WHERE {

       ?mediciones sosa:hasFeatureOfInterest onto:Ciudad_de_México.
       ?mediciones sosa:hasResult ?resultados.
       ?mediciones prov:wasAttributedTo ?Agente.
       ?resultados prov:atTime ?tiempo.

      ?mediciones geosparql:hasGeometry ?geometrias.
      ?geometrias geosparql:asWKT ?wkt.

      BIND                    (geof:buffer("<http://www.opengis.net/def/crs/OGC/1.3/CRS84>Point(-99.23
19.42)"^^<http://www.opengis.net/ont/sf#wktLiteral>, 1000, units:metre) as ?buff).

       FILTER(?tiempo > "2018-01-19T15:00:00-05:00"^^xsd:dateTime && ?tiempo < "2018-06-19T23:00:00-
05:00"^^xsd:dateTime).
      ?resultados onto:MedicionTemperatura ?temperatura

       FILTER(geof:sfContains(?buff, ?wkt)).

    }
```

Figura 5.14. Consulta para recuperar información por rango de tiempo y localización.

- Por procedencia y tipo de fuente de información.- Este filtro utiliza como parámetros de entrada el tipo de fuente de información a consultar, puede ser utilizando fuentes de datos oficiales o fuentes voluntarias, y/o la procedencia de los datos. Estas consultas se realizan por omisión sobre la Ciudad de México, obteniendo como resultado, todas las mediciones relacionadas tanto a su proveedor como al tipo de fuente de información a la que pertenece. En la Figura 5.15 se muestra la consulta para recuperar información por procedencia y tipo de fuente de información.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
    PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
    PREFIX geo: <http://www.opengis.net/ont/geosparql#>
    PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
    PREFIX gml: <http://www.opengis.net/ont/gml#>
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    PREFIX par: <http://parliament.semwebcentral.org/parliament#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    PREFIX sf: <http://www.opengis.net/ont/sf#>
    PREFIX time: <http://www.w3.org/2006/time#>
    PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
    PREFIX xml: <http://www.w3.org/XML/1998/namespace>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX prov: <http://www.w3.org/ns/prov#>
    PREFIX sosa: <http://www.w3.org/ns/sosa/>
    PREFIX onto: <http://datos.climaticosCiudades.mx/ontologia/>
    PREFIX my:<http://datos.itinerariosTuristicos.mx/recursos/>
    PREFIX geome:<http://datos.climaticosCiudades.mx/recursos/geometry/>
    PREFIX geosparql:<http://www.opengis.net/ont/geosparql#>

     SELECT DISTINCT *
     WHERE {

        ?mediciones sosa:hasFeatureOfInterest onto:Ciudad_de_México.
        ?mediciones sosa:hasResult ?resultados.
        ?mediciones prov:wasAttributedTo ?Agente.
        ?resultados prov:atTime ?tiempo.

FILTER(?tiempo > "2018-01-19T15:00:00-05:00"^^xsd:dateTime && ?tiempo < "2018-06-19T23:00:00-
05:00"^^xsd:dateTime).
       ?resultados onto:MedicionTemperatura ?temperatura

         FILTER(?Agente != "REDMET").
    }
```

Figura 5.15. Consulta para recuperar información por procedencia y tipo de fuente de información.

Una vez recuperada la información, es procesada y asignada a un modelo de medidas meteorológicas RDD (*Resilient Distributed Datasets*) y de calidad del aire. El modelo está definido por las variables meteorológicas, el proveedor de los servicios y las mediciones de calidad del aire. Estos modelos son utilizados como entrada para las diferentes operaciones a realizar.

5.5 Operaciones y experimentos

En esta sección se muestran algunos ejemplos utilizando las operaciones descritas en el capítulo 4.

5.5.1 Predicción de calidad del aire

La primera operación es la predicción de variables de calidad del aire a partir de las mediciones que se recuperan desde el *SPARQL Endpoint,* el cual contiene el repositorio de datos integrados semánticamente. A continuación, se explicará paso a paso la operación.

Primero, se realiza una consulta al *SPARQL Endpoint* (ver Figura 5.16), con el propósito de recuperar todos los datos meteorológicos y de calidad del aire de la Ciudad_de_México de ambas fuentes (voluntarias y oficiales). De esta manera, se obtiene el segundo conjunto de datos de entrenamiento, dicho conjunto de datos se muestra en la Figura 5.17.

```
PREFIX afn: <http://jena.hpl.hp.com/ARQ/function#>
    PREFIX fn: <http://www.w3.org/2005/xpath-functions#>
    PREFIX geo: <http://www.opengis.net/ont/geosparql#>
    PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
    PREFIX gml: <http://www.opengis.net/ont/gml#>
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    PREFIX par: <http://parliament.semwebcentral.org/parliament#>
    PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    PREFIX sf: <http://www.opengis.net/ont/sf#>
    PREFIX time: <http://www.w3.org/2006/time#>
    PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
    PREFIX xml: <http://www.w3.org/XML/1998/namespace>
    PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
    PREFIX rdfs: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
    PREFIX prov: <http://www.w3.org/ns/prov#>
    PREFIX sosa: <http://www.w3.org/ns/sosa/>
    PREFIX onto: <http://datos.climaticosCiudades.mx/ontologia/>
    PREFIX my:<http://datos.itinerariosTuristicos.mx/recursos/>
    PREFIX geome:<http://datos.climaticosCiudades.mx/recursos/geometry/>
    PREFIX geosparql:<http://www.opengis.net/ont/geosparql#>

     SELECT DISTINCT *
     WHERE {

       ?mediciones sosa:hasFeatureOfInterest onto:Ciudad_de_México.
       ?mediciones sosa:hasResult ?resultados.
       ?mediciones prov:wasAttributedTo ?Agente.
       ?resultados prov:atTime ?tiempo.
       ?resultados onto:MedicionTemperatura ?temperatura.
       ?resultados onto:MedicionHumedad ?humedad .
       ?resultados onto:MedicionViento ?vviento .
}
```

Figura 5.16. Consulta para recuperar todas las instancias del *SPARQL Endpoint*.

```
<?xml version="1.0"?>
<sparql xmlns="http://www.w3.org/2005/sparql-results#">
  <head>
    <variable name="mediciones"/>
    <variable name="resultados"/>
    <variable name="Agente"/>
    <variable name="tiempo"/>
    <variable name="temperatura"/>
    <variable name="humedad"/>
    <variable name="vviento"/>
  </head>
  <results>
    <result>
      <binding name="mediciones">
        <uri>http://datos.climaticosCiudades.mx/recurso/Medicion_5</uri>
      </binding>
      <binding name="resultados">
        <uri>http://datos.climaticosCiudades.mx/ontologia/Valores_5</uri>
      </binding>
      <binding name="Agente">
        <uri>http://datos.climaticosCiudades.mx/ontologia/wu-ICIUDADD45</uri>
      </binding>
      <binding name="tiempo">
        <literal datatype="http://www.w3.org/2001/XMLSchema#dateTime">2011-02-24T16:12:33-05:00</literal>
      </binding>
      <binding name="temperatura">
        <literal datatype="http://www.w3.org/2001/XMLSchema#double">26.3</literal>
      </binding>
      <binding name="humedad">
        <literal datatype="http://www.w3.org/2001/XMLSchema#double">33.0</literal>
      </binding>
      <binding name="vviento">
        <literal datatype="http://www.w3.org/2001/XMLSchema#double">4.0</literal>
      </binding>
    </result>
  </results>
</sparql>
```

Figura 5.17 Ejemplo de instancia recuperada por la consulta.

Una vez recuperada la respuesta de la consulta al *SPARQL Endpoint*, se realiza el proceso de transformación de la respuesta al modelo RDD, sustituyendo los valores que son vacíos o cero por nulos. Posteriormente, se inicializa un contexto de la biblioteca H2O y se indica que cargue el RDD a un H2O *Frame*, encargado de dividir la tabla en *n* segmento, tal como se presenta en la Figura 5.18.

```
val bigTable = spark.sql(
    """SELECT
      |ca.Fecha,ca.Hora,ca.NO_Ozono,
      |ca.NO_Azufre,ca.NO_Nitrogeno,ca.NO_CO2,
      |ca.NO_PM10,ca.NE_Ozono,ca.NE_Azufre,
      |ca.NE_Nitrogeno,ca.NE_CO2,
      |ca.NE_PM10,ca.C_Ozono,ca.C_Azufre,
      |ca.C_Nitrogeno,ca.C_CO2,ca.C_PM10,
      |ca.SO_Ozono,ca.SO_Azufre,ca.SO_Nitrogeno,
      |ca.SO_CO2,ca.SO_PM10,ca.SE_Ozono,
      |ca.SE_Azufre,ca.SE_Nitrogeno,ca.SE_CO2,
      |ca.SE_PM10,hm.Estacion,hm.Humedad,
      |hm.Temperatura, hm.VelocidadViento
      |FROM Calidaddelairedia ca
      |INNER JOIN HistoricoVariablesClima hm
      |ON ca.Fecha=hm.Fecha_nom AND ca.Hora=hm.Hora_nom
      |WHERE ca.Fecha IS NOT NULL AND ca.Hora IS NOT NULL AND ca.NO_Ozono IS NOT NULL AND
      |ca.NO_Azufre IS NOT NULL AND ca.NO_Nitrogeno IS NOT NULL AND ca.NO_CO2 IS NOT NULL AND
      |ca.NO_PM10 IS NOT NULL AND ca.NE_Ozono IS NOT NULL AND ca.NE_Azufre IS NOT NULL AND
      |ca.NE_Nitrogeno IS NOT NULL AND ca.NE_CO2 IS NOT NULL AND
      |ca.NE_PM10 IS NOT NULL AND ca.C_Ozono IS NOT NULL AND ca.C_Azufre IS NOT NULL AND
      |ca.C_Nitrogeno IS NOT NULL AND ca.C_CO2 IS NOT NULL AND ca.C_PM10 IS NOT NULL AND
      |ca.SO_Ozono IS NOT NULL AND ca.SO_Azufre IS NOT NULL AND ca.SO_Nitrogeno IS NOT NULL AND
      |ca.SO_CO2 IS NOT NULL AND ca.SO_PM10 IS NOT NULL AND ca.SE_Ozono IS NOT NULL AND
      |ca.SE_Azufre IS NOT NULL AND ca.SE_Nitrogeno IS NOT NULL AND ca.SE_CO2 IS NOT NULL AND
      |ca.SE_PM10 IS NOT NULL AND hm.Estacion IS NOT NULL AND hm.Humedad IS NOT NULL AND
      |hm.Temperatura IS NOT NULL AND  hm.VelocidadViento IS NOT NULL
      """.stripMargin)


   val train: H2OFrame = bigTable.repartition(4)
```

Figura 5.18. Código para seleccionar la información desde el RDD.

Este último conjunto de datos lo utilizamos para entrenar el algoritmo de *Deep Learning* de la biblioteca *H2O*. De acuerdo a la documentación de *H2O Deep Learning*[35] el algoritmo está basado en una red neuronal progresiva, entrenada con un gradiente descendente usando *back-propagation.* Se utiliza la configuración por omisión, (ver la Figura 5.19) donde:

- *Train* .- Es nuestro conjunto de datos recuperados.
- *Epoch* .- Son el número de iteraciones que se harán sobre los datos.
- *Activation* .-Se relaciona a la función de activación que usaremos (Tahn, Tahn with dropout, Rectifier, Rectifier with dropout, Maxout, Maxout with dropout)
- *Hidden* .- Se refiere al número de capas ocultas.

[35] http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html

```
val dlParams = new DeepLearningParameters()
   dlParams._train = train
   dlParams._epochs = 5
   dlParams._activation = Activation.RectifierWithDropout
   dlParams._hidden = Array[Int](100, 200,100)

   val dl = new DeepLearning(dlParams)
   val dlModel = dl.trainModel.get
```

Figura 5.19 Parámetros de configuración del algoritmo de *Deep Learning*.

Por último, se utiliza un arreglo unidimensional de variables como el mostrado en la tabla 5.3:

Tabla 5.3. Vector de ejemplo.

Fecha	Hora	Estación	Humedad	Temperatura	Velocidad
01/03/11	12	V1	50	24	10

El vector contiene la fecha y la hora como datos obligatorios, los cuales se utilizan para generar la predicción, este vector se introduce al algoritmo previamente entrenado y regresa la siguiente respuesta en un archivo de texto (ver Figura 5.20).

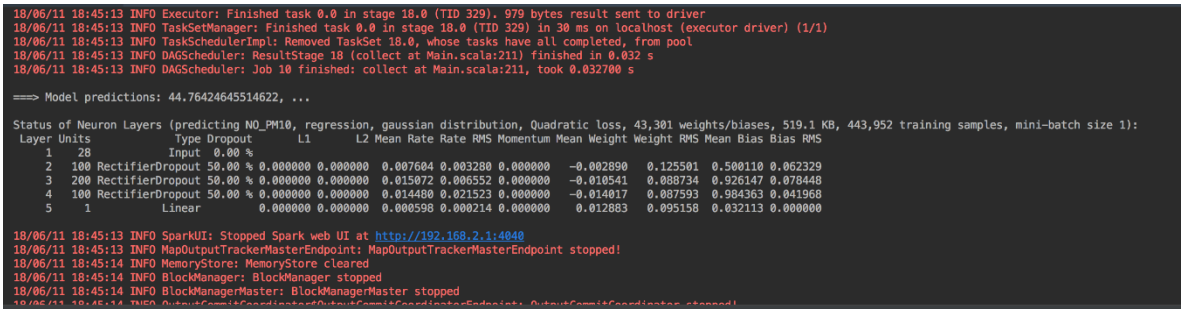


Figura 5.20. Resulta de predicción de la primera variable NO_PM10.

Acorde al análisis, podemos comparar la respuesta conforme a los datos oficiales, por lo tanto, el valor oficial que arroja el histórico es de 47, el obtenido con las fuentes VGI son 44.76, teniendo una diferencia de 2.24. Además, de acuerdo a los resultados y utilizando las tablas de rangos definidas por el Sistema de Monitoreo Atmosférico, es posible obtener la categoría en la que se define la calidad del aire, como se muestra en la Figura 5.21.

Tabla 3. Equivalencias para Partículas Menores a 10 micrómetros (PM_{10}).

Concentración de PM_{10} (Promedio móvil de 24h)	Concentraciones para los puntos de corte ($\mu g/m^3$)		Equivalencia en el índice para los puntos de corte		k	Categoría
	BP_{Hi}	BP_{Lo}	I_{Hi}	I_{Lo}		
0 - 40	40	0	50	0	1.2500	BUENA
41 - 75	75	41	100	51	1.4412	REGULAR
76 - 214	214	76	150	101	0.3551	MALA
215 - 354	354	215	200	151	0.3525	MUY MALA
355 - 424	424	355	300	201	1.4348	EXTREMADAMENTE MALA
425 - 504	504	425	400	301	1.2532	
505 - 604	604	505	500	401	1.0000	

Figura 5.21. Tabla de equivalencia para el PM10.

5.5.2 Comparativa de medidas VGI

La siguiente operación a realizar es la comparativa entre cada uno de los orígenes de las medidas VGI contra las medidas oficiales, verificando si existen discrepancias entre las medidas recolectadas y las proporcionadas por fuentes gubernamentales. Se utilizan las medidas oficiales como la línea base (Gold Standard), debido a que están proporcionadas por entidades expertas, en particular haremos una consulta que nos regrese todas las medidas oficiales del 19 de abril del 2011 en un horario de las 15:00 a las 22:00 de fuentes oficiales, como se ve en la Tabla 5.4.

Tabla 5.4. Resultado parcial de la consulta de fuentes oficiales al *SPARQL Endpoint*.

Fecha	Hora	Estación	Humedad	Temperatura	Velocidad
19/04/11	15	CUA	30	22.5	1.4
19/04/11	15	FAC	18	27.9	1.7
19/04/11	15	SAG	16	28	0.3
19/04/11	15	SUR	34	21.6	4.2
19/04/11	15	TLA	18	28.1	1
19/04/11	15	TPN	54	16.2	6.2
19/04/11	15	VIF	20	25	1.6
19/04/11	15	XAL	15	27.4	0.7

De la anterior consulta se recuperan en total 195 mediciones de 10 posibles estaciones meteorológicas y de calidad de aire que no tuvieran ningún valor vacío o igual a cero. Posteriormente se procesa y se transforma la respuesta en un modelo RDD para su comparación. Como segundo conjunto de datos se realiza la misma consulta, pero únicamente recuperando las fuentes de información voluntarias, el resultado se puede ver en la Tabla 5.5.

Tabla 5.5. Resultado parcial de la consulta de fuentes VGI.

Fecha	Hora	Estación	Temperatura
19/04/11	15	aerisapi	32
19/04/11	16	wu-ICIUDADD45	24.2
19/04/11	16	aerisapi	27
19/04/11	16	wu-ICIUDADD21	18.2
19/04/11	17	wu-ICIUDADD45	17
19/04/11	17	wu-ICIUDADD21	18.8
19/04/11	18	aerisapi	32
19/04/11	18	wu-ICIUDADD45	24.2

En total se recuperan 35 instancias de la consulta que no tuvieron ningún valor vacío o igual a cero, de la misma forma con la que se obtuvo el conjunto de datos oficiales, éste también se procesa y se transforma en un modelo RDD. Para la comparativa entre las medidas se calcula el error que tienen las mediciones voluntarias con respecto a las mediciones oficiales con base en las estaciones meteorológicas que se ubiquen dentro de un radio de 10 kilómetros, de igual manera se utilizan como parámetros de comparación la fecha y hora de cada medición como se ven en la Tabla 5.6 y la Figura 5.22.

Tabla 5.6 Errores en las mediciones voluntarias

Hora	Estación oficial	Temp oficial (ºC)	Estación VGI	Temp VGI (ºC)	Porcentaje de Error
15	UIZ	25.3	aerisapi	32	20.9375
16	UIZ	23	wu-ICIUDADD45	24.2	4.958677686
16	UIZ	23	aerisapi	27	14.81481481
16	UAX	14.5	wu-ICIUDADD21	18.2	20.32967033
17	UIZ	19.1	wu-ICIUDADD45	17	-12.35294118
17	UAX	18.4	wu-ICIUDADD21	18.8	2.127659574
18	UIZ	16.5	wu-ICIUDADD45	15.8	-4.430379747
18	UAX	16.4	wu-ICIUDADD21	17	3.529411765
19	UIZ	17	wu-ICIUDADD45	16.3	-4.294478528
19	UIZ	17	aerisapi	19	10.52631579
19	UAX	15.8	wu-ICIUDADD21	14.5	-8.965517241
20	UIZ	17.4	wu-ICIUDADD45	17	-2.352941176
20	UIZ	17.4	aerisapi	18.7	6.951871658
20	UAX	14.7	wu-ICIUDADD21	14	-5
21	UIZ	17.2	wu-ICIUDADD45	16.5	-4.242424242
21	UIZ	17.2	aerisapi	18	4.444444444
21	UAX	14.2	wu-ICIUDADD21	13.6	-4.411764706
22	UIZ	16.8	wu-ICIUDADD45	16	-5
22	UIZ	16.8	aerisapi	17.4	3.448275862

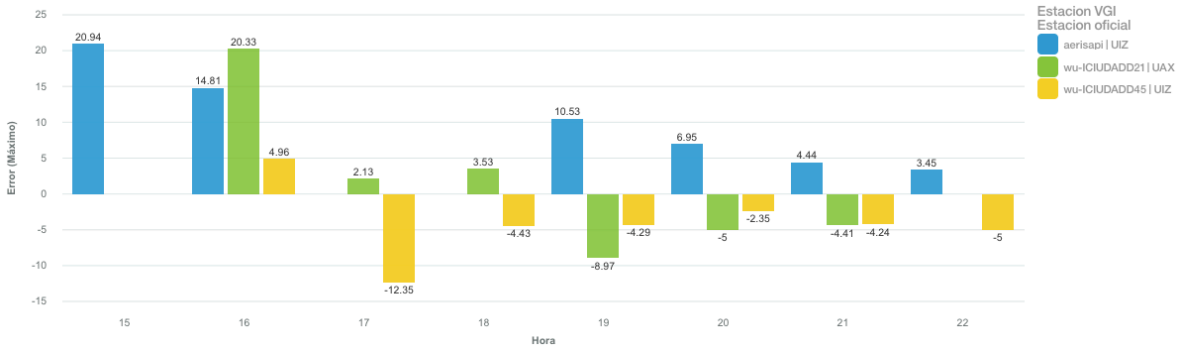


Figura 5.22. Porcentaje de error de las medidas voluntarias por hora.

De tal manera que podemos determinar el error promedio durante el período de tiempo sobre el que se hizo la consulta, dando como resultado que la estación meteorológica wu-ICIUDADD21 que pertenece al agente *WeatherUnderground*, como estación meteorológica de menor error con respecto a las otras tres, como se ve en la Figura 5.23.

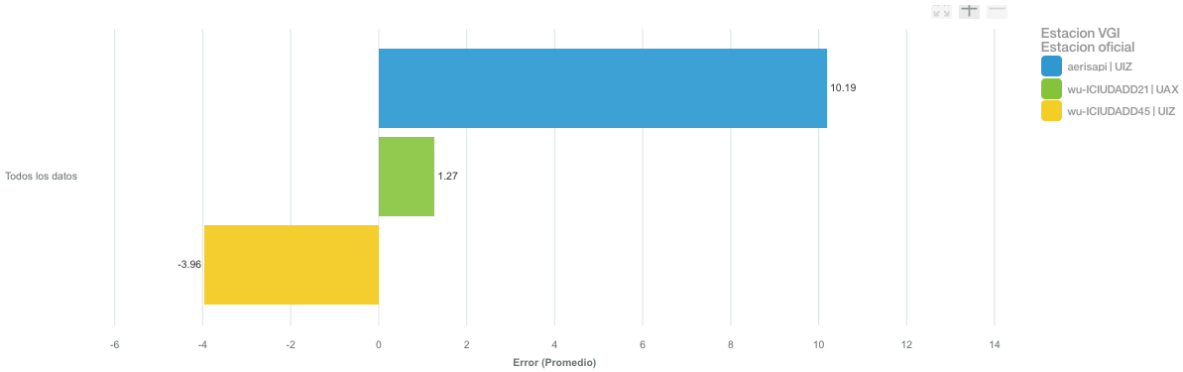


Figura 5.23. Error promedio por estación meteorológica.

De igual modo, en la Tabla 5.7 podemos ver la cantidad de medidas con error dentro del rango de incertidumbre del 10% con respecto a la medida oficial, dándonos como resultado que la estación meteorológica wu-ICIUDAD45 del agente *WeatherUnderground* ofrece mediciones más precisas con respecto a las medidas oficiales.

Tabla 5.7 Mediciones correctas y erróneas por estación meteorológica.

Estación VGI	Medidas correctas	Medidas incorrectas	Porcentaje de medidas correctas
aerisapi	3	3	50%
wu-ICIUDADD45	6	1	85.71%
wu-ICIUDADD21	5	1	83.33%

Para validar la respuesta del sistema, utilizaremos las medidas *Precision*, *Recall* y *F1*, estas medidas son comunes en la recuperación de información. Estas medidas están basadas en la comparación de un resultado esperado y en la efectividad de la respuesta del sistema de evaluación, los resultados son considerados un conjunto de *ítems* (Euzenat, 2007), en nuestro caso un conjunto de medidas meteorológicas y de calidad del aire.

El *Precision* y el *Recall* estas definidos por las siguientes formulas:

$$\boldsymbol{Precision} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FP}} \quad (1)$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FN}} \quad (2)$$

$$\boldsymbol{F} = \mathbf{2} * \frac{\boldsymbol{Precision*Recall}}{\boldsymbol{Precision+Recall}} \quad (3)$$

Donde las variables están definidas de la siguiente manera:

- Verdaderos positivos (VP).- Son representados por todas las mediciones comparadas que están dentro de la medida estándar.
- Falsos positivos(FP).- Son todas aquellas mediciones que fueron encontradas, pero no están dentro de la medida estándar.
- Falsos negativos(FN).- Representan a todos las mediciones que no fueron encontradas y están dentro de la medida estándar.
- Verdaderos negativos(VN).- Las medidas que están en esta categoría son las que no pertenecen a la medida estándar y no fueron encontradas.

Para nuestro caso de uso, definiremos los conceptos de la siguiente forma:

- Verdaderos positivos(VP).- Son representados como las medidas correctas que están dentro del rango de incertidumbre del gold standard.
- Falsos positivos(FP).- Representan a todas las mediciones que están fueran del rango de incertidumbre del gold standard.

- Falsos negativos(FN).- Son todas aquellas mediciones que no tienen correspondencia con alguna medida del gold standard.

- Verdaderos negativos(VN).- Son aquellas medidas que no existen tanto en el conjunto que se compara como en las medidas del gold standard.

A partir de la Tabla 5.6 se obtienen los valores requeridos para calcular el *Precision* y *Recall* como se presenta en la Tabla 5.8.

Tabla 5.8. Evaluación de cada fuente de información voluntaria.

Fuente de información	VP	FP	VN	FN	Precision	Recall	F
Aeris weather	3	3	0	2	0.50	0.60	0.5454
wu-ICIUDADD45	6	1	0	1	0.85	0.85	0.85
wu-ICIUDADD21	5	1	0	2	0.83	0.71	0.7653

Como se puede observar la fuente de información “Aeris weather” requiere de mejoras en su funcionamiento y muestreo de información ya que su medida F es la más baja con 54.54%. Caso contrario de la fuente de información “wu-ICIUDADD45” del agente “WeatherUnderground” que presenta un comportamiento más estable y tiene la mejor medida F con 85%, por lo tanto, se concluye que sería la mejor fuente de información voluntaria con la que se cuenta hasta el momento.

CAPÍTULO 6 CONCLUCIONES Y TRABAJO A FUTURO

Durante el desarrollo de este trabajo se diseño y desarrollo una metodología capaz de integrar de manera semántica diversas fuentes de información, mediante la implementación de los cuatro componentes propuestos en la metodología. Para el proceso de la recolección de datos se diseño e implemento un programa el cual este encargado de recuperar la información de las fuentes de datos, para las oficiales(estáticas) se hace mediante la lectura de los archivos CSV, para las voluntarias(dinámicas) se realiza mediante consultas programadas hacia sus servicios web o APIs, al final el programa manda la información a un *stream* de datos para su posterior procesamiento.

Después se realiza un pre procesamiento de la información donde se recupera la información que viaja en el *stream* de datos y se generan los diferentes tipos de objetos conforme a los formatos recuperados(XML, JSON, CSV), sobre los cuales se realiza una búsqueda de las variables de meteorológicas y de calidad del aire empleando el archivo de sinónimos, al final se genera un objeto que contenga los valores de las mediciones, origen y localización que será procesado por el siguiente módulo.

Para el proceso de integración semántica de las fuentes de información se diseñó e implementó una red ontológica utilizando diversos escenarios de la metodología NeOn. La red de ontologías cuenta con la incorporación de diversos estándares internacionales como GeoSPARQL para el modelado de la información geográfica, la SSN y su núcleo SOSA para el modelado de los sensores, sus procesos y mediciones y la Prov-o, para el modelado del origen, producción y consumo de la información. La red de ontologías construida brinda una estructura semántica necesaria para el proceso de integración de las diferentes variables asociadas con fuentes de información meteorológica y de calidad del aire.

Finalmente, la información integrada semánticamente resultante genera un repositorio semántico de datos sobre el cual se puede realizar una explotación con diversos propósitos, este repositorio tiene la capacidad de realizar operaciones espaciales sobre la información, gracias al estándar de GeoSPARQL.

La metodología propuesta, aporta soluciones a las limitantes encontradas durante el desarrollo del trabajo, como la falta de canales o *streamings* de datos que brindan la información, lo que genera que la información no se trate en tiempo real, para esto se implemento el uso de la herramienta Apache Kafka que nos permite crear y consumir *streamings* de datos, además se realizó el diseño e implementación del módulo de recopilación de información encargado de hacer consultas a diversas fuentes de información dinámicas y estáticas.

El otro aporte fue la solución a la cantidad de diferentes etiquetas que recibe una sola variable debido a que no existe un compendio de términos que tenga todas las etiquetas relacionadas con una palabra. La solución encontrada fue que al momento de recuperar las respuestas y agregar al diccionario de sinónimos las diferentes etiquetas que fuimos encontrando conforme se fueron agregando diferentes fuentes de información voluntaria.

En cuanto a los alcances, se logró integrar las diversas fuentes de datos en un repositorio de datos semánticos, siendo enriquecido con una variedad de información a parte de los valores de mediciones meteorológicas, como el origen de los datos y su proveedor, su posición geográfica (si se recoge como información en la fuente) , además de la aplicación de los principios de *Linked Data* para generar y publicar información acerca de las ciudades que se están observando y de las propiedades que se observan.

En cuanto a las operaciones que se pueden realizar, existen varios filtros, uno de ellos es por localización para esto se emplearon operaciones espaciales que se realizan en el *SPARQL EndPoint*, esto fue posible por la importación de la

ontología de *GeoSPARQL* que le permite a las instancias tener una componente espacial con la que se pueden realizar operaciones como *contains* y *buffer,* que se utilizan para realizar los filtrados de las consulta*s,* entre otras que se pueden realizar.

Acorde con los objetivos particulares descritos en la sección 1.4, se logro su cumplimiento durante el desarrollo de este trabajo como se ve a continuación. El primer objetivo se cumple con un programa para recuperar la información de las diferentes fuentes de información, tanto estáticas como dinámicas como se ve en la sección 5.4.1. El segundo objetivo se realiza mediante el diseño y desarrollo módulo de pre-procesamiento que se describe en la sección 5.4.2. El tercer objetivo se lleva acabo mediante el diseño y la implementación de la red de ontologías, además la integración semántica se describe en la sección 5.4.3. El último objetivo se refleja en el diseño del módulo encargado de la comparación de calidad que se ve reflejado en la implementación del componente de análisis en la sección 5.4.4.

6.1 Trabajo a Futuro

Como líneas de trabajo futuro y con el objetivo de dar continuidad a este trabajo se propone abordar los siguientes tópicos; La implementación de un *SPARQL Streaming*, que permita realizar consultas sobre un *streaming* de datos, generado a partir del repositorio semántico de datos, esto permitirá que se empleen herramientas como C-SPARQL que proporcionan la capacidad de realizar las consultas sobre un *stream* de información mediante el registro de consultas con ventanas de tiempo; Implementar un algoritmo de autodescubrimiento que permita incorporar nuevas fuentes de información, así como la integración y completitud de nuevas variables meteorológicas (por ejemplo: dirección del viento, Presión, etc.).y de calidad del aire (por ejemplo: PM2,5, hidrocarburos totales (HTC), Hidrocarburos No Metánicos (HCNM),entre otros) .

BIBLIOGRAFÍA

Ahmad, S., & Simonovic, S. P. (2006). An intelligent decision support system for management of floods. Water Resources Management, 20(3), 391-410.

Ahmada, M., Alia, A., & Khiyalb, M. S. H. (2016). Potential of Volunteered Geographic Information for Adaptation of Climate Change Effects in Pakistan.

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. Applied Geography, 49, 45-53.

Apache Spark. (2016). Apache spark. Retrieved from http://spark.apache.org/

Bacon T (2013) Big bang? When 'Big Data' gets too Big. http://www.eyefortravel.com/mobile-and-technology/big-bang-when-%E2%80%98big-data%E2%80%99-gets-too-big. Visitado el 16 Abril 2018

Bhattacharjee, S., Das, M., Ghosh, S. K., & Shekhar, S. (2016, October). Prediction of meteorological parameters: an a-posteriori probabilistic semantic kriging approach. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (p. 38). ACM.

Blázquez, L. M. V., Gargantilla, J. Á. R., Corcho, O., & i Subirana, J. C. (2008). Hacia una armonización semántica de la información geográfica. Treballs de la Societat Catalana de Geografia, 727-736.

Blázquez, L. M. V., Pascual, A. F. R., Ángel, M., Poveda, B., & de Aplicaciones Geográficas, S. G. (2006). Ingeniería ontológica: El camino hacia la mejora del acceso a la información geográfica en el entorno web. *Subdirección General de*

Aplicaciones Geográficas del Instituto Geográfico Nacional. Avances En Las Infraestructuras De Datos Espaciales, 95.

Bendre, M. R., Thool, R. C., & Thool, V. R. (2015, September). Big data in precision agriculture: Weather forecasting for future farming. In Next Generation Computing Technologies (NGCT), 2015 1st International Conference on (pp. 744-750). IEEE.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, *5*(3), 1-22.

de Brito Moreira, R., Degrossi, L. C., & de Albuquerque, J. P. (2015, April). An experimental evaluation of a crowdsourcing-based approach for flood risk management. In 12th Workshop on Experimental Software Engineering (ESELAW), At Lima, Peru (pp. 1-11).

Brody, T. A., Flores, J., French, J. B., Mello, P. A., Pandey, A., & Wong, S. S. (1981). Random-matrix physics: spectrum and strength fluctuations. *Reviews of Modern Physics*, *53*(3), 385.

Buriano, L.; Marchetti, M.; Carmagnola, F.; Cena, F.; Gena, C.; Torre, I., (10-12 May 2006) *The Role of Ontologies in Context-Aware Recommender Systems*, Mobile Data Management, 2006. MDM 2006. 7th International Conference on (pp.80).

C. Allocca, M. D'Aquin, and E. Motta. DOOR - Towards a Formalization of Ontology Relations. In Jan L. G. Dietz, editor, KEOD, pages 13–20. INSTICC Press, 2009

Campbell, A. T., Eisenman, S. B., Lane, N. D., Miluzzo, E., Peterson, R. A., Lu, H., ... & Ahn, G. S. (2008). The rise of people-centric sensing. IEEE Internet Computing, 12(4).

Candillier, L.; Meyer, F.; Boullé, M.; (2007) *Comparing state-of-the-art collaborative filtering systems*, Lecture Notes in Computer Sciece, 4571, (pp. 548–562).

Castro Degrossi, L., Porto de Albuquerque, J., Restrepo-Estrada, C. E., Mobasheri, A., & Zipf, A. (2017). Exploring the geographical context for quality assessment of VGI in flood management domain.

Chapman, L., Bell, C., & Bell, S. (2017). Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. International Journal of Climatology, 37(9), 3597-3605.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly*, 1165-1188.

Cressie, N., & Wikle, C. K. (2015). Statistics for spatio-temporal data. John Wiley & Sons.

Corcho, Ó.; Fernández, M.; Gómez, A.; López, A., (2005) *Construcción de ontologías legales con la metodología METHONTOLOGY y la herramienta WebODE*.

Cruz, A. P. H., Ortega, C. A. P., & Granados, L. A. P. (2017). DISEÑO DE UNA ONTOLOGÍA PARA AGENTES QUE MONITOREAN MEDICIONES DE SENSORES. REVISTA COLOMBIANA DE TECNOLOGIAS DE AVANZADA (RCTA), 2(26).

Degrossi, L. C., de Albuquerque, J. P., Fava, M. C., & Mendiondo, E. M. (2014). Flood Citizen Observatory: a crowdsourcing-based approach for flood risk management in Brazil. In SEKE (pp. 570-575).

Demchenko, Y., De Laat, C., & Membrey, P. (2014, May). Defining architecture components of the Big Data Ecosystem. In *Collaboration Technologies and Systems (CTS), 2014 International Conference on* (pp. 104-112). IEEE.

D'Hondt, E., Stevens, M., & Jacobs, A. (2013). Participatory noise mapping works! An evaluation of participatory sensing as an alternative to standard techniques for environmental monitoring. Pervasive and Mobile Computing, 9(5), 681-694.

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier.

Duhigg C (2012) How companies learn your secrets. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html. Visitado el 16 Mayo 2018

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. Innovation Policy and the Economy, 14(1), 1-24.

Euzenat, J. (2007, January). Semantic Precision and Recall for Ontology Alignment Evaluation. In *IJCAI* (Vol. 7, p. 348353).

Fernández, M.; Gómez, A.; Juristo, N., (1997) *METHONTOLOGY: From Ontological Art Towards Ontological Engineering*, AAAI Symposium on Ontological Engineering, Stanford.

Fischer, F. (2012). VGI as big data. *GeoInformatics*, *3*, 46-47.

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. GeoJournal, 72(3-4), 137-148.

Frandsen, M., Paton, D., & Sakariassen, K. (2011). Fostering community bushfire preparedness through engagement and empowerment. Australian Journal of Emergency Management, The, 26(2), 23.

Garrido J. y I. Requena. (2010). Gestión de conocimiento aplicado a evaluación de impacto ambiental mediante ontologías. Revista Ambientalia, 1: 129-140.

Gibbons, P. B., Karp, B., Ke, Y., Nath, S., & Seshan, S. (2003). Irisnet: An architecture for a worldwide sensor web. IEEE pervasive computing, 2(4), 22-33.

Ghodsi, M. (2014). A brief review of recent data mining applications in the energy industry. International Journal of Energy and Statistics, 2(01), 49-57.

Gibbons, P. B., Karp, B., Ke, Y., Nath, S., & Seshan, S. (2003). Irisnet: An architecture for a worldwide sensor web. IEEE pervasive computing, 2(4), 22-33.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211-221.

Godchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110-120.

Gorman, S. (2012). Why VGI is the Wrong Acronym, fortius One.

Grosky, W. I., Kansal, A., Nath, S., Liu, J., & Zhao, F. (2007). Senseweb: An infrastructure for shared sensing. IEEE multimedia, 14(4).

Gruber, T., (1995) *Towards principles for the design of ontologies used for knowledge sharing*, International Journal of Human-Computer Studies, 43(5/6), (pp. 907-928).

Guarino, N., (1995) *Formal ontology, conceptual analysis and knowledge representation*, International Journal of Human-Computer Studies, 43(5- 6), (pp. 625-640).

Guzmán, J. A.; López, M.; Torres, I. D., (Mayo 2012) *Metodologías y métodos para la construcción de ontologías*. Scientia et Technica, [S.l.], 2(50), (pp. 133-140).

Hadoop, A. (2014). Welcome to Apache Hadoop. Retrieved from http://haddop.apache.org

Hamm S (2013) How big data can boost weather forecasting. http://readwrite.com/2013/02/28/how-big-data-can-boost-weather-forecasting#awesm=ou64ZEaKe2HtUu. Visitado el 20 Mayo 2018

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and technique

Hand, D. J. (2009). Mining the past to determine the future: Problems and possibilities. International Journal of Forecasting, 25(3), 441-451.

Hartig, O., Bizer, C., & Freytag, J. C. (2009, October). Executing SPARQL queries over the web of linked data. In *International Semantic Web Conference* (pp. 293-309). Springer, Berlin, Heidelberg.

Harvey, F. (2013). To volunteer or to contribute locational information? Towards truth in labeling for crowdsourced geographic information. In Crowdsourcing Geographic Knowledge (pp. 31-42). Springer, Dordrecht.

Hassani, H., & Silva, E. S. (2015). Forecasting with big data: A review. Annals of Data Science, 2(1), 5-19.

Haupt, S. E., & Kosovic, B. (2015, December). Big data and machine learning for applied weather forecasts: Forecasting solar power for utility operations. In Computational Intelligence, 2015 IEEE Symposium Series on (pp. 496-501). IEEE.

Haworth, B., & Bruce, E. (2015). A review of volunteered geographic information for disaster management. Geography Compass, 9(5), 237-250.

Haworth, B., Bruce, E., & Middleton, P. (2015). Emerging technologies for risk reduction: assessing the potential use of social media and VGI for increasing community engagement. Australian Journal of Emergency Management, The, 30(3), 36.

Hengl, T., Heuvelink, G. B., Tadić, M. P., & Pebesma, E. J. (2012). Spatio-temporal prediction of daily temperatures using time-series of MODIS LST images. Theoretical and applied climatology, 107(1-2), 265-277.

Hjort, J., Suomi, J., & Käyhkö, J. (2011). Spatial prediction of urban–rural temperatures using statistical methods. Theoretical and applied climatology, 106(1-2), 139-152.

Horita, F. E., de Albuquerque, J. P., Degrossi, L. C., Mendiondo, E. M., & Ueyama, J. (2015). Development of a spatial decision support system for flood risk management in Brazil that combines volunteered geographic information with wireless sensor networks. Computers & Geosciences, 80, 84-94.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... & Twigger, S. (2008). Big data: The future of biocuration. *Nature*, *455*(7209), 47.

Hubbard, D. W. (2014). How to measure anything: Finding the value of intangibles in business. John Wiley & Sons.

Hung, K. C., Kalantari, M., & Rajabifard, A. (2016). Methods for assessing the credibility of volunteered geographic information in flood response: A case study in Brisbane, Australia. Applied Geography, 68, 37-47.

Intergovernmental Panel on Climate Change (IPCC) 2012, Managing the risks of extreme events and disasters to advance climate change adaptation, In Field CB, Barros V, Stocker TF, Qin D, Dokken DJ, Ebi KL, Mastrandrea MD, Mach KJ, Plattner G-K, Allen SK, Tignor M & Midgley PM (eds), A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge and New York.

Ježek, J., Jedlička, K., & Martološ, J. (2015). Visual Analytics of Traffic-Related Open Data and VGI. In ICIST 2015 Conference (pp. 13-26).

Knapp A (2013) Forecasting the weather with big data and the fourth dimension. http://www.forbes.com/sites/alexknapp/2013/06/13/forecasting-the-weather-with-big-data-and-the-fourth-dimension/2/. Visitado el 20 Mayo 2018

Kilibarda, M., Hengl, T., Heuvelink, G., Gräler, B., Pebesma, E., Perčec Tadić, M., & Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. Journal of Geophysical Research: Atmospheres, 119(5), 2294-2313.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy, 52(6), 119-139.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, 21(1), 1-24.

Laloux, L., Cizeau, P., Potters, M., & Bouchaud, J. P. (2000). Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, *3*(03), 391-397.

Lanfranchi, V., Wrigley, S. N., Ireson, N., Wehn, U., & Ciravegna, F. (2014, January). Citizens' observatories for situation awareness in flooding. In ISCRAM 2014 Conference Proceedings-11th International Conference on Information Systems for Crisis Response and Management (pp. 145-154). Sheffield.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. IEEE Communications magazine, 48(9).

Larin-Fonseca, R., & Garea-Llano, E. (2013). Método de Enriquecido Semántico para la Integración de Objetos Geoespaciales. Ciencias de la Tierra y el Espacio, 14(1), 60-69.

Li, X., Lin, F., & Qiu, R. C. (2014). Modeling massive amount of experimental data with large random matrices in a real-time UWB-MIMO system. *arXiv preprint arXiv:1404.4078*.

Longley, P. (2005). Geographic information systems and science. John Wiley & Sons.

Longueville, B. D., Luraschi, G., Smits, P., Peedell, S., & Groeve, T. D. (2010). Citizens as sensors for natural hazards: A VGI integration workflow. Geomatica, 64(1), 41-59.

Ludwig, T., Reuter, C., & Pipek, V. (2015). Social haystack: dynamic quality assessment of citizen-generated content during emergencies. ACM Transactions on Computer-Human Interaction (TOCHI), 22(4), 17.

Madden, S. (2012). From databases to big data. IEEE Internet Computing, 16(3), 4-6.

Mazzoleni, M., Verlaan, M., Alfonso, L., Monego, M., Norbiato, D., Ferri, M., & Solomatine, D. P. (2017). Can assimilation of crowdsourced data in hydrological modelling improve flood prediction?. Hydrology and Earth System Sciences, 21(2), 839.

Meier, F., Fenner, D., Grassmann, T., Otto, M., & Scherer, D. (2017). Crowdsourcing air temperature from citizen weather stations for urban climate research. Urban Climate, 19, 170-191.

Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., ... & Campbell, A. T. (2008, November). Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. In Proceedings of the 6th ACM conference on Embedded network sensor systems (pp. 337-350). ACM.

Montes de Oca - Pérez, Ariel; Rosario - Ferrer, Yiezenia. (2014). Ontología de evaluación de impacto ambiental para proyectos mineros. Instituto Superior Minero Metalúrgico de Moa Dr. Antonio Nuñez Jiménez' Holguín, Cuba. Revista Minería y Geología, 30 (1), 104-117.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010, November). Towards quality metrics for OpenStreetMap. In Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems (pp. 514-517). ACM.

Morales, J. J. C., Pérez, T. V., & Velásquez, A. M. P. (2017). ONTOLOGIA COMO BASE DE CONOCIMIENTO PARA LA EVALUACIÓN BIÓTICA DE LA CALIDAD DEL AGUA. REVISTA COLOMBIANA DE TECNOLOGIAS DE AVANZADA (RCTA), 1(29).

Moreno Rubio J., Jiménez López A, Barrera Lombana N. (2013). El amplificador de potencia de carga sintonizada. Revista colombiana de tecnologías de Avanzada. 2(22). Pág. 9 – 13

Neches, R., Fikes, R.E., Finin, T., Gruber, T.R., Patil, R., Senator, T. & Swartouy, W.R., (1991) *Enabling technology for knowledge sharing*, AI Magazine, 12(3), (pp. 16-36).

Ostermann, F. O., & Spinsanti, L. (2011, April). A conceptual workflow for automatically assessing the quality of volunteered geographic information for crisis management. In Proceedings of AGILE (Vol. 2011, pp. 1-6).

Paton, D. (2003). Disaster preparedness: a social-cognitive perspective. Disaster Prevention and Management: An International Journal, 12(3), 210-216.

Paradesi, S. M. (2011). Geotagging Tweets Using Their Content. In FLAIRS Conference.

Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. RFID Journal, 8(1), 204-218.

Perry, J. W., Kent, A., & Berry, M. M. (1955). Machine literature searching x. machine language; factors underlying its design and development. *Journal of the Association for Information Science and Technology*, *6*(4), 242-254.

Poser, K., & Dransch, D. (2010). Volunteered geographic information for disaster management with application to rapid flood damage estimation. Geomatica, 64(1), 89-98.

Poveda Villalon, M. (2009). Red de ontologías para el Camino de Santiago.

Prober, S., O'Connor, M., & Walsh, F. (2011). Australian Aboriginal peoples' seasonal knowledge: a potential basis for shared understanding in environmental management. Ecology and Society, 16(2).

Pyle, D. (2003). Business modeling and data mining. Morgan Kaufmann.

Qiu, R., & Wicks, M. (2014). *Cognitive networked sensing and big data*. Springer New York.

Ramos Esmeralda y Haydemar Núñez (2007). ONTOLOGÍAS: componentes, metodologías, lenguajes, herramientas y aplicaciones - Reporte Técnico: RT-2007-12. Lecturas en Ciencias de la Computación ISSN 1316- 6239. Facultad de Ciencias. Universidad Central de Venezuela. 45 p. http://lia.ciens.ucv.ve/LIA/publicaciones.php

Reddy, S., Burke, J., Estrin, D., Hansen, M., & Srivastava, M. (2007, November). A framework for data quality and feedback in participatory sensing. In Proceedings of the 5th international conference on Embedded networked sensor systems (pp. 417-418). ACM.

Rey, T., & Wells, C. (2012). Integrating data mining and forecasting-Leveraging time-series data offers the best possible forecasting model. OR/MS Today, 39(6), 34.

Richards, N. M., & King, J. H. (2013). Three paradoxes of big data. Stan. L. Rev. Online, 66, 41.

Roche, C., (2003) *Ontology : a survey*, in: 8th Symposium on Automated Systems Based on Human Skill and Knowledge, (pp. 28–41), Gteborg, Sweden.

Rogova, G. L., & Nimier, V. (2004, June). Reliability in information fusion: literature survey. In Proceedings of the seventh international conference on information fusion (Vol. 2, pp. 1158-1165).

Silver, N. (2012). The signal and the noise: the art and science of prediction. Penguin UK.

Simonovic, S. P. (1999). Decision support system for flood management in the Red River Basin. Canadian Water Resources Journal, 24(3), 203-223.

Suárez, M. C.; Gómez, A.; Fernández, M., (2012). The NeOn Methodology for Ontology Engineering. In Ontology Engineering in a Networked World, (pp. 9-34), Springer Berlin Heidelberg.

Sui, D., Elwood, S., & Goodchild, M. (Eds.). (2012). *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Springer Science & Business Media.

Sure,Y.; Staab, S.; Studer.R., (2003*) On-to-knowledge methodology*, Handbook on Ontologies, Series on Handbooks in Information Systems, 6, (pp. 117-132).

Taylor, M., Wells, G., Howell, G., & Raphael, B. (2012). The role of social media as psychological first aid as a support to community resilience building. Australian Journal of Emergency Management, The, 27(1), 20.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography, 46(sup1), 234-240.

Tran, P., Shaw, R., Chantry, G., & Norton, J. (2009). GIS and local knowledge in disaster management: a case study of flood risk mapping in Viet Nam. Disasters, 33(1), 152-169.

Tucker P (2013) The future is not a destination. http://www.slate.com/articles/technology/future_tense/2013/10/futurist_magazine_s_predictions_on_quantum_computing_big_data_and_more.html. Visitado el 2 Abril 2018

Uschold, M.; Gruninger,M., (1996) *Ontologies: principles, methods and applications*. The Knowledge Engineering Review, 11, (pp. 93-136).

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28.

Vilches-Blázquez, L. M. (2011). *Metodología para la integración basada en ontologías de información de bases de datos heterogéneas en el dominio hidrográfico* (Doctoral dissertation, Topografia).

Vilches-Blázquez, L. M., & Gargantilla, J. Á. R. (2012). Conflación semántica: un estudio sobre la integración de información geoespacial basada en ontologías. GeoFocus. Revista Internacional de Ciencia y Tecnología de la Información Geográfica, (12), 147-171.

Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O., & Gómez-Pérez, A. (2014). Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth*, *7*(7), 554-575.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hübner, S. (2001, August). Ontology-based integration of information-a survey of existing approaches. In *IJCAI-01 workshop: ontologies and information sharing* (Vol. 2001, pp. 108-117).