UNIVERSIDAD NACIONAL DE COLOMBIA

Obtener un método para la extracción de información a partir de documentos semiestructurados producidos al interior del Servicio Nacional de Aprendizaje SENA, permitiendo su publicación, reutilización e intercambio a través de la web semántica

Jhonatan Alejandro Abello Diaz

Universidad Nacional de Colombia

Facultad de Ingeniería

Bogotá D.C., Colombia

2015

Obtener un método para la extracción de información a partir de documentos semiestructurados producidos al interior del Servicio Nacional de Aprendizaje SENA, permitiendo su publicación, reutilización e intercambio a través de la web semántica

Jhonatan Alejandro Abello Diaz

Trabajo Final de Maestría presentado como requisito parcial para optar al título de:
MAGISTER EN INGENIERIA – INGENIERIA DE SISTEMAS Y COMPUTACIÓN

Director:
MSc. Ing. Mario Armando Rosero Muñoz

Línea de Investigación:
Extracción de información, Gestión del conocimiento, Web 3.0. Datos Enlazados, RDF, Ontología y web semántica.

Universidad Nacional de Colombia
Facultad de Ingeniería
Bogotá D.C., Colombia
2015

Dedicatoria

Gracias a Jehová Dios, a mis padres, mi hermano y mi esposa, ya que sin su apoyo incondicional nada de esto sería posible.

Si Jehová está por nosotros, ¿quién estará contra nosotros?

Romanos 8:31

Agradecimientos

Al Ing. Mario Armando Rosero por su paciencia, comprensión, guía y apoyo incondicional, a la Universidad Nacional de Colombia por brindarme el conocimiento y las habilidades necesarias para el desarrollo de este trabajo final de maestría. Al SENA por ser la Entidad que me ha brindado el apoyo laboral por más de 5 años.

Resumen

Actualmente en el Servicio Nacional de Aprendizaje SENA, existen gran cantidad de archivos, los cuales contienen información textual de manera semiestructurada, lo cual dificulta realizar consultas SQL complejas sobre la información allí contenida, impidiendo que esta información pueda ser utilizada de manera activa al interior de la Entidad.

Aunque actualmente la entidad posee un avanzado gestor documental, el cual se encarga de gestionar, almacenar e indexar los documentos producidos por procesos realizados al interior de la entidad, la información que se puede extraer de los mismos es bastante limitada, obligando en muchas ocasiones a abrir el documento para poder observar con mayor detalle el contenido en su interior.

Además la indexación de estos documentos, en la mayoría de los casos se realiza 100% manual, lo que expone a la entidad a errores humanos debidos a los altos volúmenes de documentos generados, así como a las múltiples fuentes que los generan; Esto impide que la información histórica contenida en estos documentos sea utilizada eficazmente como soporte en la toma de decisiones de la entidad.

Para dar una alternativa de solucion a este problema es necesario construir una base de conocimiento siguiendo la estructura y los lineamientos de datos enlazados, que permitan que esta información relevante pueda ser publicada, consultada y usada como insumo vital en la toma de decisiones al interior de la entidad.

Para esto durante el desarrollo de este trabajo se pretende obtener un método para la extracción de información a partir de documentos semiestructurados producidos al interior del Servicio Nacional de Aprendizaje SENA,

Este método será plasmado en un prototipo que permitirá extraer la información necesaria mediante cuatro fases que abarcan desde la Extracción de Información hasta la fase de Persistencia de conocimiento, de manera que sea posible inferir la información requerida.

Palabras clave: Extracción de información, Gestión del conocimiento, Datos Enlazados, RDF, Ontología, Web Semántica, SENA.

Abstract

Now in the Servicio Nacional de Aprendizaje SENA, there are lots of files, which contain textual semi-structured information, making it difficult to perform complex SQL queries about the information contained therein, preventing this information can be actively used inside SENA.

Although the company now has an advanced document management system, which is responsible for managing, storing and indexing the documents produced by processes performed inside SENA, the information can be extracted from them is very limited, forcing many times to open the document to observe in detail the contents inside.

Moreover indexing of these documents, in most cases 100% manually, which exposes the entity to human error due to high volumes of documents generated, as well as multiple sources that generate performed, this prevents the historical information contained in these documents to be used effectively as a support in the decision making in the organization.

To give an alternative solution to this problem is necessary to build a knowledge base following the structure and guidelines linked data, which allow this relevant information can be posted, accessed and used as vital input in decision making inside the entity.

For this during the development of this work it is to obtain a method for extracting information from semi-structured documents produced inside SENA,

This method is embodied in a prototype which will extract the necessary information through four stages ranging from extraction to the phase information persistence of knowledge, so that it is possible to infer the required information

Keywords: Information Extraction, Knowledge Management, Linked Data, RDF, Ontology, Semantic Web, SENA

Contenido

Lista de figuras

Introducción

En el Servicio Nacional de Aprendizaje SENA, existen gran cantidad de archivos, los cuales contienen información textual de manera semiestructurada, lo cual dificulta realizar consultas SQL complejas sobre la información allí contenida, de manera que dificulta que esta información pueda ser utilizada constantemente al interior de la Entidad.

Actualmente la entidad posee un gestor documental, el cual apoya el proceso de gestion, almacenamiento e indexación de los documentos producidos al interior de la entidad, sin embargo la información disponible relacionada con estos documentos es bastante limitada.

Además la indexación de estos documentos, en la mayoría de los casos se realiza 100% manual, lo cual expone a la entidad a errores humanos debidos a los altos volúmenes de documentos generados, así como a las múltiples áreas al interior de la Entidad que los generan, dificultando que la información histórica contenida en estos documentos sea utilizada eficazmente como soporte en la toma de decisiones de la entidad, realización de prospección e inteligencia de negocios, los cuales son prácticas vitales en cualquier entidad.

Para lo tanto con el objetivo de mitigar este problema, es necesario construir una base de conocimiento que se ajuste a la estructura y los lineamientos de datos enlazados, de manera que se permita que esta información r pueda ser publicada, consultada y usada como insumo vital en la toma de decisiones al interior de la entidad.

Por lo tanto para dar continuidad al desarrollo de este trabajo final de maestría, es indispensable identificar el contexto del problema planteado y así poder definir los objetivos propuestos a dar cumplimiento durante el desarrollo del mismo.

1. Descripción General del Proyecto

En el método propuesto en este trabajo, aborda el problema de extracción de información para el poblado de una ontología destinada a la Web Semántica. Esta extracción de datos se realiza a partir de documentos de texto semiestructurado producidos al interior del Servicio Nacional de Aprendizaje SENA, con la finalidad de poblar un modelo semántico. La tarea de extraer segmentos de datos que representen la semántica subyacente en una ontología es de alta complejidad, por lo tanto es preciso someter a prueba el método propuesto en este trabajo.

Una de las estrategias más ampliamente acertadas para evaluar una representación semántica es mediante la formulación y aplicación de reglas de inferencia. También con el fin de evaluar la capacidad de razonamiento y extracción de información se formularon consultas que están destinadas a recuperar información producida por estos procesos.

Estas consultas tienen dos objetivos:

- Evaluar la precisión de la inferencia de información.
- Validar la correcta extracción de información.

1.1 Objetivo General

Implementar un prototipo del método obtenido para la extracción de información textual a partir de documentos semiestructurados generados al interior del Servicio Nacional de Aprendizaje SENA, facilitando la publicación, consulta e intercambio de información contenida en estos documentos.

1.2 Objetivos Específicos

- Analizar y Evaluar los métodos y técnicas existentes utilizadas para la extracción de información textual semiestructurada.
- Definir un conjunto de reglas para la extracción y manipulación de información textual semiestructurada.
- Definir una estructura de datos semántica, a partir de la información obtenida de los documentos semiestructurados.
- Diseñar un esquema para la obtención de conocimiento, a partir de la información contenida en documentos de texto semiestructurado.
- Implementar un prototipo del modelo obtenido en un escenario real al interior de la entidad; este modelo debe permitir evaluar los métodos y técnicas existentes, para la definición, unificación, transformación, publicación, enlace y generación de información.
- Obtener un método para representar la semántica de la información contenida en los documentos de texto semiestructurados producidos al interior de la entidad.
- Validar si realmente los datos obtenidos con éxito realmente representan la información original (contenida en los documentos de texto semiestructurado).
- Recopilar las experiencias y resultados obtenidos al implementar el prototipo desarrollado, para la extracción de información textual a partir de documentos semiestructurados generados al interior del Servicio Nacional de Aprendizaje SENA

2. Marco Teórico

El marco teórico contiene algunos conceptos esenciales, los cuales permitan una mayor entendimiento del desarrollo del trabajo final de maestría, también se presentaran algunas investigaciones previas en áreas afines al tema, así como algunos avances relacionados encontrados en la literatura existente.

2.1 Marco Conceptual

A continuación se incluyen algunos conceptos esenciales, los cuales son necesarios para la compresión y un mejor entendimiento del trabajo desarrollado en los capítulos siguientes.

2.1.1 Extracción de Información

La Extracción de Información (EI) es la disciplina, que pretende mediante diversos algoritmos, métodos, procesos y herramientas, identificar fragmentos de información o subcadenas de caracteres que representan hechos o permitan inferir algún hecho relevante, de manera que la semántica del contenido de los documentos (Lavelli et al., 2008) pueda ser representada.

Es decir permite alimentar una base de conocimiento, representando los hechos como instancias de los respectivos conceptos representados.

El campo de la Extracción de Información (EI) fue impulsado por el programa MUC (Message Understanding Conference) de DARPA (Defense Advanced Research Projects Agency) en 1987 (Grishman & Sundheim, 1996).Allí precisamente se definió a la EI como la tarea de extraer hechos específicos y bien definidos desde texto proveniente de conjuntos de documentos homogéneos en un dominio restringido, a este dominio se le llama dominio de extracción (Jurafsky & Martin, 2008) y rellenando los campos predefinidos en plantillas con la información extraída.

El programa MUC ha inspirado una gran cantidad de trabajos en EI y se ha convertido en una referencia importante en el campo de la minería de texto; pese a esto aún es bastante complejo diseñar e implementar herramientas de extracción de información que gocen de altos niveles de precisión y confiablidad.

Un sistema de Extracción de Información (EI) diseñado para el procesamiento de texto consta típicamente de 3 pasos (Hobbs et al., 1997):

Preprocesamiento de Texto: Se realiza segmentación de información, la cual varía desde fragmentos de texto simple inmersos en oraciones y/o frases, hasta fragmentos de palabras inmersas en oraciones y/o frases complejas, con el objetivo de realizar un análisis lingüístico completo.

Selección de Reglas: Suelen definirse palabras claves, las cuales buscan permitir recorrer el texto para identificar el elemento clave, de esta manera las reglas correspondientes se activan y permiten extraer el patrón de texto encontrado.

Aplicación de Reglas: Se selecciona la regla que cumple con las condiciones predefinidas y así permite rellenar una plantilla con la información textual extraída.

Típicamente las reglas son declaradas a priori, estas normalmente están basadas en estructuras lógicas, es decir en forma de expresiones regulares y/o patrones. Realmente lo que se busca es alimentar un campo de una plantilla determinada. De esta manera si la información estructurada almacenada en la plantilla se ha extraído correctamente, puede ser transformada y utilizada como insumo para alguna base de conocimiento (ontología); así queda totalmente demostrado que los recursos de texto son fuentes extremadamente útiles de conocimiento para diseñar, modelar, alimentar y enriquecer ontologías.

La EI puede clasificarse de acuerdo a diferentes criterios (Chen Shao-fei, Tian-zhu, & Yang Wen-zhu, 2003) Según:

1. **Tipo de Recurso:**
    - Texto plano.
    - Texto semiestructurado.
    - texto estructurado.

2. **Método de Extracción:**
    - Método de extracción basado en procesamiento de lenguaje natural.
    - Método basado en ontologías.
    - Métodos de aprendizaje automático.
    - Métodos basados en la estructura de HTML o XML

2.1.2 Expresiones Regulares (Regex)

Las expresiones regulares en el desarrollo de software son un método el cual permite realizar búsquedas dentro de cadenas de caracteres (Thompson, 1968), sin importar la cantidad de caracteres de la búsqueda, encontrando todas las apariciones de un patrón previamente definido de caracteres en un documento textual. Adicionalmente la búsqueda de patrones permite validar formatos específicos en una cadena de caracteres específica, como puede ser fechas, identificadores, acentuación entre otros.

Las expresiones regulares (ER) son un recurso muy utilizado por gran cantidad de programas en los que se hace imprescindible la búsqueda y reemplazo de información (Junghoo Cho, 2012), En otras palabras son un lenguaje para definir exactamente lo que se necesita buscar. Es común que las ER se utilicen conjuntamente con otros recursos como: la detección de entidades y las técnicas de recuperación de información para la extracción de información.

Para el caso puntual de este trabajo final de maestría, se pretende generar el prototipo de una herramienta que sea capaz de completar información de forma automática, es decir sin necesidad de teclear nuevamente información o por lo menos que este proceso sea reducido de manera significativa.

2.1.3 PDF/A

El PDF/A (PDF/A Competence Center, 2011) Es un subconjunto de PDF el cual fue obtenido excluyendo características sobrantes para el archivado a largo plazo. Este estándar establece una serie de requisitos a los programas para la visualización de archivos PDF/A y está definido por la ISO 19005-1:2005.1.

Este estándar define unas especificaciones para los documentos electrónicos, de manera que se garantice que estos puedan ser reproducidos a futuro con exactitud. La principal clave para lograr este objetivo es que los documentos en formato PDF/A deben estar 100% auto-contenidos, es decir que toda la información necesaria para mostrar de forma consistente el documento deberá estar presente en el archivo. Lo cual incluye: el contenido (texto, imágenes, gráficos), las fuentes utilizadas e inclusive la información del color. Esto implica que en un documento PDF/A no está permitido:

- Depender de fuentes externas (p.ej. programas de tipografía o hiperenlaces).
- Uso de audio y de vídeo.
- Uso de JavaScript o la ejecución de archivos.
- Todas las fuentes deben estar integradas y sin restricciones
- Los espacios de color deben estar especificados de manera independiente al dispositivo.
- No está permitido el uso de cifrado.
- El empleo de metadatos basados en estándares es obligatorio

2.1.4 Web Semántica

La Web Semántica (McIlraith, Plexousakis, & Harmelen, 2004) es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida.

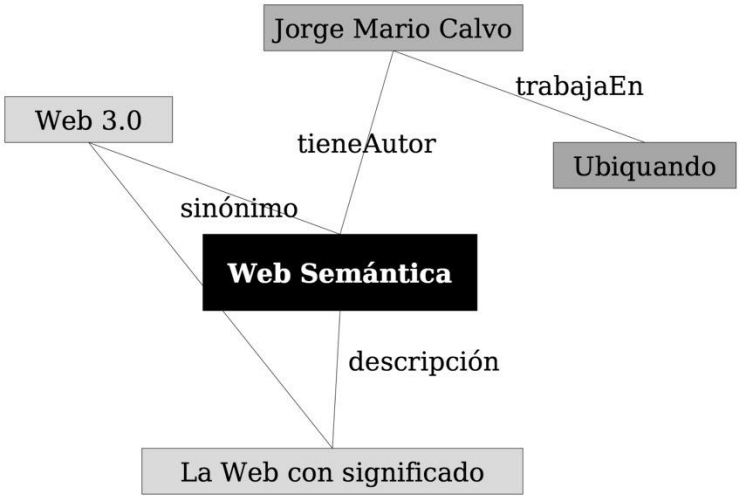


Figura 1. Web Semántica (Calvo L., 2006)

Dota a la Web de más significado y, por lo tanto, de más semántica, se pueden obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común, mediante la cual, es posible compartir, procesar y transferir información de forma sencilla. Esta Web extendida y basada en el significado, se apoya en lenguajes universales que resuelven los problemas ocasionados por una Web carente de semántica en la que, en ocasiones, el acceso a la información se convierte en una tarea difícil y frustrante.

En resumen, la web semántica es una extensión de la actual web cuyo objetivo es que no solo los humanos, sino también las maquinas, sean capaces de “comprender” el contenido de los documentos (Peis, Herrera-Viedma, Hassan, & Herrera, 2003)

2.1.5 Componentes de la Web Semántica

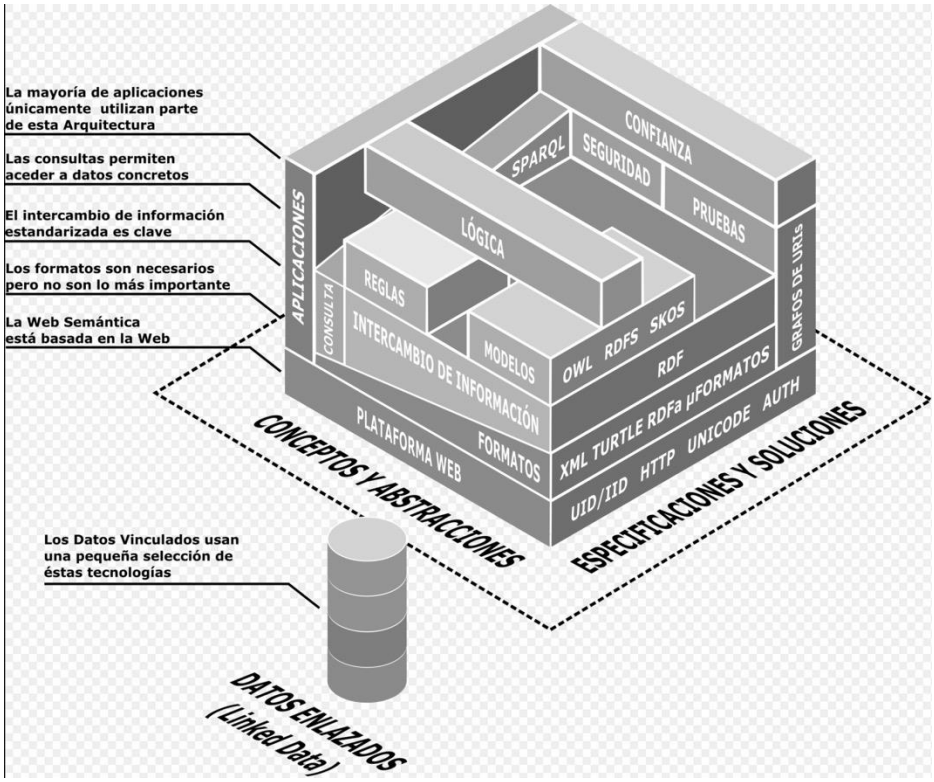


Figura 2. Arquitectura Tecnológica de la Web Semántica (Arroyo & Otros, 2008)

Los principales componentes de la Web Semántica son los metalenguajes y los estándares de representación XML, XML Schema, RDF, RDF Schema y OWL, así como el lenguaje SPARQL para la consulta de datos RDF (Juan Antonio Pastor Sánchez, 2011; Rodríguez Mendez, 1999). La OWL (Owl Web Ontology Language Overview) describe la función y relación de cada uno de estos componentes de la Web Semántica (Dimitrova et al., 2008):

- **XML** (BOSAK & Tim Bray, 1999): Aporta la sintaxis superficial para los documentos estructurados, pero sin dotarles de ninguna restricción sobre su significado.

- **XML Schema** (Altova, 2007): Es un lenguaje para definir la estructura de los documentos XML

- **RDF** (Rodríguez Mendez, 1999): Es un modelo de datos para los recursos y las relaciones que se puedan establecer entre ellos. Aporta una semántica básica para este modelo de datos que puede representarse mediante XML.

- **RDF Schema** (W3C, 2004): Es un vocabulario para describir las propiedades y las clases de los recursos RDF, con una semántica para establecer jerarquías de generalización entre dichas propiedades y clases.

- **OWL** (Rodríguez Mendez, 1999): Es un lenguaje para definir ontologías mediante la descripción detallada de propiedades y clases: tales como relaciones entre clases (p.ej. disyunción), cardinalidad (por ejemplo "únicamente uno"), igualdad, tipologías de propiedades más complejas, caracterización de propiedades (por ejemplo simetría) o clases enumeradas.

- **SPARQL** (Angles & Gutierrez, 2008; Buil-Aranda, Hogan, Umbrich, & Vandenbussche, 2013): Es un lenguaje de consulta de conjuntos de datos RDF. Además en dicha especificación también se incluye un formato XML que detalla el modo en el que se estructuran los resultados obtenidos.

2.1.6 Linked Data (Datos Enlazados)

Los Datos Enlazados (Desire Consortium, 2000; Piedra, N., Tovar, E., Colomo-Palacios, R., Lopez-Vargas, J., & Chicaiza, J.A., 2014), es la forma que tiene la Web Semántica de vincular o relacionar los 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 (McIlraith et al., 2004) no se trata únicamente de la publicación 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, pudiendo llegar a información relacionada que se hace referencia desde otros datos iniciales.

Es decir esta alternativa está destinada tanto para usuarios técnicos como para agentes de software inteligentes, debido a que estas dos entidades pueden comprender la semántica del dominio representado mediante el modelo y con ello realizar diferentes procesos que permitan obtener información particular de individuos que no se encuentren representados completamente en el modelo particular; También es necesario la utilización de un y un lenguaje de consulta semántico (SPARQL).

Linked Data define 4 principios (W3C Working Group, 2006):

- Los recursos de información deben de ser nombrados con URIs.
- Uso de URIs http.
- Instalación de la infraestructura de software para la consulta de información.
- Inclusión de enlaces RDF (Antoine Isaac, 2009).

2.1.7 Marco de Trabajo Para Descripción de Recursos

El Marco de trabajo para descripción de recursos (Resource Description Framework - RDF) fue creado en 1997 bajo el auspicio de la W3C (W3C Working Group, 2006), con el fin de tener un formato de compatibilidad entre los diversos sistemas de metadatos, suministrando para ello una arquitectura genérica de metainformación. Por ese motivo la W3C decidió utilizar el lenguaje XML como sistema de comunicación.

2.1.8 Formato RDF

El formato RDF tiene como origen dos ramas modernas de la Documentación (Johan Hjelm, 2001). Por una parte los metadatos (los cuales permiten interconectar sistemas entre sí) y, por otra parte la representación del conocimiento a través del concepto de Web Semántica.

RDF posee la capacidad de representar metadatos y así facilitar la interoperabilidad entre diversas aplicaciones, brindando un mecanismo para el intercambio de información a través de la Web. El formato RDF tiene distintas áreas de aplicación como: catalogación del material bibliotecario digital, la recuperación de recursos de información, agentes

inteligentes, sistemas de gestión de propiedad intelectual, entre otras aplicaciones.

El principal objetivo de RDF (World Wide Web Consortium, 2012) es definir un mecanismo, el cual permita describir recursos que tengan como principios la independencia de plataforma y la interoperabilidad de metadatos, de manera que permita fusionar distintas descripciones de recursos realizados con diferentes conjuntos de metadatos (Ora Lassila, 1997).

El bloque básico de construcción en RDF es una tripleta conformada por un sujeto, predicado y objeto, comúnmente denotada con P(O,S). Esto es, un sujeto tiene un predicado (o propiedad) con un valor:

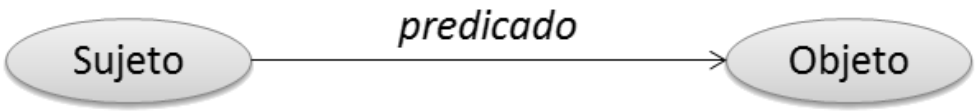


Figura 3. Ejemplo del bloque básico de construcción de una tripleta RDF

Esta anotación es de bastante utilidad, permitiendo que sujetos y objetos puedan ser intercambiados. Así, cualquier objeto de una tripleta puede tomar el papel de un sujeto en otra tripleta.

Las principales características de RDF son:

- **Neutral:** Al no estar ligado explícitamente a ningún otro sistema de metadatos ni plataforma.
- **Expresivo:** Las etiquetas utilizadas son intuitivas.
- **Procesable:** Al ser texto ASCII con una estructura bien definida.

Por estos motivos han aparecido en los últimos años gran cantidad de herramientas que facilitan la representación de metadatos mediante RDF. En la siguiente sección se mencionan algunas herramientas utilizadas para el manejo de metadatos en RDF.

2.1.8.1 Librerías y Herramientas RDF

La posibilidad de contar con herramientas que tengan la capacidad de gestionar y procesar información en RDF es vital, tanto para crear y/o consultar información en este formato. Por esto es recomendable utilizar una API que se encargue de las tareas básicas de procesamiento de RDF.

Sin embargo el principal problema surge en la amplia variedad de alternativas ya que cada API ofrece mejores rendimientos que otras o, incluso, ofrecen una mayor cantidad de características para el procesamiento y gestión de datos en RDF, además de estar implementadas en diferentes lenguajes de programación, como Java, PHP, Perl, Python, Entre las más utilizadas podemos encontrar las siguientes:

- **RDF API para PHP (RAP):** Es un conjunto de herramientas para la Web Semántica dirigida a desarrolladores en PHP.

- **RDFLIB:** Es una librería que permite manipular grafos en RDF en lenguaje Python. RDFLib cuenta con un parser/serializador RDF/XML que cumple con la especificación de la sintaxis RDF/XML. Esta librería permite almacenar en base de datos relacionales como MySQL, SQLite, Berkeley DB entre otras, además cuenta con soporte para consultas en SPARQL.

- **SESAME (Daniel Krech, 2012):** Es un framework para el procesamiento de datos en RDF. Sesame incluye un analizador sintáctico, manejo de persistencia, inferencia y consultas de datos en RDF.

  Además ofrece una API la cual permite conectarse a las principales herramientas almacenamiento de RDF.

- **JENA (Broekstra, Kampman, & Harmelen, 2002):** Es un framework de Java utilizado para el desarrollo de aplicaciones para la Web Semántica, el cual brinda una amplia gama de librerías escritas en Java para RDF, RDFS, RDFa, OWL y SPARQL.

También incluye un motor de inferencia el cual permite realizar razonamiento basado en ontologías OWL y RDFS.

- **PROTÉGÉ** (“The Protege Project”, 2000)**:** Es un editor de ontologías y sistema de adquisición de conocimientos de código abierto. Protege también proporciona una interfaz gráfica de usuario para definir ontologías.

  Cabe destacar que en muchas de estas aplicaciones se utiliza RDF Schema para modelar el dominio de conocimiento mediante la declaración de conceptos (clases), propiedades y sus interrelaciones.

2.1.9 RDF Schema

RDF Schema es un mecanismo que permite a los desarrolladores definir un vocabulario para la representación de datos en RDF y así especificar el tipo de objetos para los cuales los predicados pueden ser aplicados. RDF-S o RDFS realiza esto mediante la preespecificación de alguna terminología como Class, subClassOf y Property, la cual puede ser usada en esquemas de aplicación específica. Las expresiones RDFS son también expresiones RDF válidas.

La propiedad subClassOf permite representar la organización jerárquica de clases, Además los objetos pueden ser declarados como instancias de estas clases usando la propiedad type; También permite declarar restricciones para el uso de propiedades mediante la especificaciones de Rango y Dominio.

Para entender mejor el significado semántico de la terminología presentada en RDFS, por encima de línea punteada de la imagen 4, se ilustra el RDFS que define el vocabulario utilizado en la expresión RDF. Persona, Contrato y Supervisor son declarados como Class (clases) y supervisa es declarada como una propiedad, por debajo de la línea punteada se declaran instancias/objetos mediante los términos de este vocabulario.

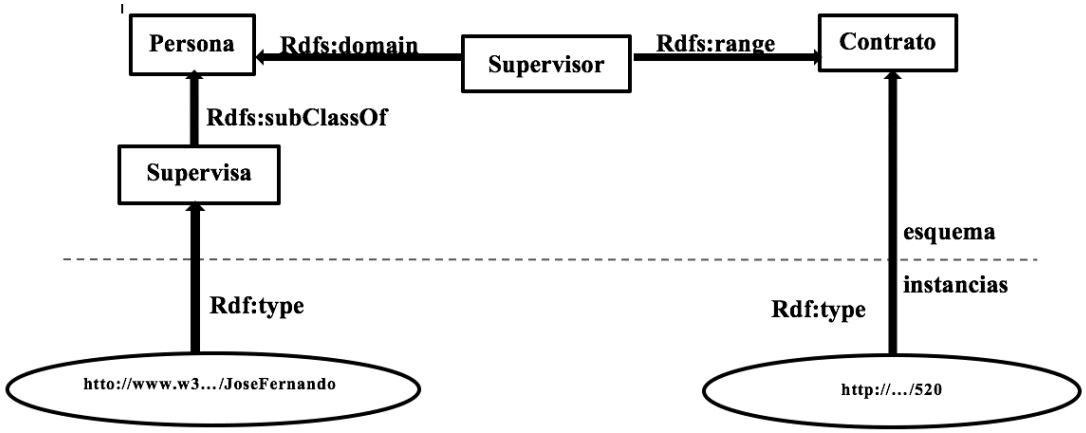


Figura 4. Ejemplo ilustrado del esquema y las instancias de un grafo RDF

En este trabajo para lograr la representación de conocimiento se usara RDF y RDFS ya que presentan un modelo de datos extensamente usado en la Web Semántica para la representación e inferencia de conocimiento.

2.1.10 Ontologías

La ontología es la rama de la filosofía, las ontologías estudian la naturaleza del ser y la existencia. El término ontología se deriva de la palabra griega onto, la cual significa ser y logia que significa discurso escrito o hablado (Blackwell, 2004). En ciencias computacionales se define una nueva interpretación de ontología como: Una especificación explícita de una conceptualización compartida (Gruber, 1993). En la Web Semántica la reutilización del conocimiento es facilitada por uso de ontologías explícitas, es decir, donde el conocimiento es codificado en alguna forma de representación formal. Por lo tanto, se hace necesario definir un lenguaje para ontologías con tres aspectos importantes.

- **Conceptualización:** El lenguaje debe elegir un modelo de referencia adecuado. Además, debe proporcionar las primitivas correspondientes para representar conocimiento, tales como la definición de entidades y relaciones en un dominio.

- **Vocabulario:** Además de la semántica el lenguaje también debe tener en cuenta la sintaxis, así como la simbología apropiada para indicar los conceptos, así como la gramática para la representación de la conceptualización en una representación explícita.

- **Axiomatización:** Con el fin de capturar la semántica para la inferencia de nuevo conocimiento, es necesaria la implementación de reglas y restricciones, además del conocimiento de hechos.

También se debe garantizar que los agentes de software necesarios para compartir conocimiento por medio de la web, estén de acuerdo sobre ontologías comunes y así permitir el intercambio de conocimiento y cumplir con las metas de colaboración. Para compartir conocimiento a través de diferentes dominios se deben tener presentes tres aspectos al desarrollar una ontología:

- **Extensibilidad:** La ontologías se debe desarrollar de manera incremental, reutilizando la mayor cantidad de conceptos como sea posible antes de crear un nuevo concepto desde cero.

- **Visibilidad:** El hecho de publicar conocimiento en la Web, no garantiza que éste pueda ser realmente entendido por máquinas y/o humanos. Con el fin de que el conocimiento sea visible en la Web se requieren acuerdos en común sobre la sintaxis y la semántica entre el productor y el consumidor, ya que los agentes de software no son capaces de entender el conocimiento escrito en un lenguaje desconocido.

- **Inferencia:** Una ontología no sólo sirve para la representación del conocimiento de un dominio en particular, sino que también sirve para propósitos de cálculo, ya que permite la inferencia lógica de hechos mediante la axiomatización.

De manera que es necesario que las ontologías en la Web proporcionen las estructuras necesarias para enlazarlas con las primitivas de inferencia lógica y sean capaces de soportar una variedad de requerimientos de complejidad y expresividad.

Precisamente en los últimos años se han desarrollado una amplia variedad de lenguajes para la creación de ontologías que permiten representar el conocimiento con la suficiente expresividad y formalidad para que sea capaz de ser comprensible por herramientas de software y así cubrir las características que se mencionaron anteriormente.

2.1.11 Lenguajes Para la Construcción de Ontologías

En los últimos años varios lenguajes han sido desarrollados para representación del conocimiento mediante ontologías. Precisamente la construcción de estos lenguajes está evolucionando de acuerdo a un enfoque en capas (capas de razonamiento e inferencia).

Ahora bien, es importante que estos lenguajes cumplan un determinado número de requerimientos para ser útiles para el modelado de sistemas inteligentes. Estos requerimientos son:

- **Sintaxis Razonable.**
- **Semántica Bien Definida.**
- **Capacidad de Expresividad**.
- **Razonamiento Eficiente y Comprensible.**
- **Útil Para Construir Grandes Bases de Conocimiento.**

Por lo tanto un lenguaje ontológico debe describir el significado de una manera que sea comprensible para las computadoras, es decir un lenguaje para construir ontologías no sólo necesita tener la habilidad para definir el vocabulario, sino también los medios para definir formalmente la manera en la que trabajará el razonamiento automático. Actualmente existe un gran énfasis en los lenguajes de ontologías web. Dado que la Web está descentralizada, los lenguajes deben permitir definir diversos vocabularios distribuidos.

Actualmente las ontologías se han convertido en tema de investigación en diversas disciplinas de la computación. Por lo tanto, las ontologías están siendo desarrolladas e implementadas en distintos campos de la ciencia con sus propias teorías, formalismos, enfoques y particularidades.

Es necesario tener presente que el principal propósito de las ontologías no es solo servir como vocabularios y taxonomías, sino que debe permitir compartir y reutilizar conocimiento a través de agentes inteligentes y aplicaciones, ya que una ontología de cierto dominio tanto su terminología (vocabulario del dominio), los conceptos esenciales en el dominio, su clasificación, taxonomía, relaciones (incluyendo todas las jerarquías importadas y sus restricciones) y axiomas del dominio.

En la Web Semántica las ontologías juegan un rol vital en ayudar a los procesos automatizados (llamados agentes inteligentes) para acceder a la información. Debido a que las ontologías proporcionan un número de características útiles para los sistemas inteligentes, así como también para la representación del conocimiento en general.

La gran ventaja de las ontologías es que ofrecen un vocabulario bien estructurado que explica las relaciones entre diferentes términos, permitiendo a los agentes inteligentes interpretar su significado de manera flexible pero sin ambigüedades. Por lo tanto, una ontología proporciona un vocabulario y un entendimiento común el cual es procesable por computadoras o herramientas de software, para esto es necesario conocer cuál es y cómo se describe el lenguaje para la especificación de ontologías utilizado en la Web Semántica, OWL.

2.1.12 Lenguajes de Ontologías Web

El lenguaje de ontologías web (Web Ontology Language - OWL) es un lenguaje que permite definir y crear instancias de ontologías web. Las ontologías OWL pueden incluir definiciones de clases, propiedades y sus instancias. La semántica formal de OWL especifica cómo derivar sus consecuencias lógicas, es decir, hechos no representados explícitamente en la ontología, pero que implica la semántica.

Estas implicaciones se pueden basar en un simple documento o múltiples documentos distribuidos, los cuales deben ser combinados usando mecanismos definidos en OWL. Los términos cuyo significado es definido en ontologías pueden ser usados en un marco

semántico que describe el contenido y funcionalidad de los recursos web disponibles (Berners-Lee, 2000).

OWL toma la habilidad de enunciar hechos en RDF y las capacidades de estructurar clases y propiedades de RDF Schema. OWL permite declarar clases y organizarlas en una jerarquía de tipo subclase_de, como con RDFS. Las clases OWL pueden ser especificadas como combinaciones lógicas (intersecciones, uniones o complementos) de otras clases o como la enumeración de objetos específicos. También permite declarar propiedades, organizar éstas en una jerarquía de propiedades como subpropiedad_de e indicar el rango y dominio de aplicación de estas propiedades, como con RDFS. El dominio de las propiedades OWL son las clases, el rango puede ser clases o tipos de datos definidos como una cadena o enteros.

Con OWL se puede establecer que una propiedad es transitiva, simétrica, funcional o es inversa de otra propiedad; con esto OWL extiende a RDFS. De manera que OWL puede expresar que objetos/individuos pertenecen a cierta clase y qué valores de las propiedades son de cierto tipo de individuos. También permite declarar clases equivalentes y propiedades, clases disjuntas, entre otras.

Sin embargo el mayor aporte de OWL al extender a RDFS es la habilidad para indicar restricciones sobre cómo se comportan las propiedades que son locales a una clase. De manera que se pueden definir clases donde una propiedad en particular es restringida para que todos los valores sean de cierta clase (o un tipo de dato) o al menos un valor debe ser de cierta clase.

2.1.12.1 OWL Lite

Es la variante más sencilla para la creación de una ontología en términos de clases, propiedades, relaciones de subclases y restricciones. OWL Lite es una fracción de OWL DL con el objetivo de proporcionar un subconjunto útil de características del lenguaje. OWL Lite no incluye las sintaxis para crear uniones, complementos e individuos en las descripciones o axiomas de clase. Además, limita su cardinalidad a 0 o 1.

OWL Lite es usada comúnmente para representar taxonomías, por ejemplo existe (Kriglstein & Wallner, 2010) una herramienta para la visualización de ontologías OWL-Lite, esta herramienta está destinada para ontologías con una jerarquía extendida y gran número de instancias.

2.1.12.2 OWL DL

OWL DL es una variante sintáctica de la Lógica Descriptiva SHOIN (Horrocks, Patel-Schneider, & Van Harmelen, 2003).OWL DL ofrece mayor expresividad pero garantizando completitud computacional, es decir todas las conclusiones son computables y decidibles.

Esto quiere decir que su lógica descriptiva ofrece constructores adicionales, tales como conjunción, disyunción y negación, además de clases y relaciones. También OWL DL posee dos importantes mecanismos de inferencia: subsunción, es decir incluir un elemento en una clasificación más general, y consistencia, es decir, verificar que se cumplan las restricciones establecidas.

2.1.12.3 OWL FULL

OWL FULL ofrece la máxima expresividad y libertad sintáctica pero sin garantías computacionales. OWL FULL Permite aumentar el significado del vocabulario predefinido (en RDF o en OWL), sin embargo actualmente ningún software de razonamiento es capaz de soportar razonamiento completo para cualquier característica de OWL FULL.

La mayor diferencia entre OWL-DL y OWL FULL es que el espacio de clases y el espacio de instancias son disjuntos en OWL-DL pero no en OWL FULL. Esto quiere decir que una clase puede ser interpretada como un conjunto de individuos y como un individuo perteneciente a otra clase en OWL-FULL. El vocabulario completo de OWL puede ser utilizado sin ningún tipo de restricciones en OWL-FULL (Djeddi & Khadir, 2010; Khadir, Djeddai, & Djeddi, 2011).

Para el desarrollo del caso de estudio de este trabajo, el modelo de ontologías OWL fue seleccionado para representación del conocimiento, debido a su extenso uso en la Web Semántica y a la gran variedad de herramientas para el desarrollo de ontologías OWL. Además, tiene características idóneas para la representación semántica de información

en nuestro caso de prueba.

2.1.13 Lenguajes de Consulta

Varios lenguajes formales han sido desarrollados para consultar información en grafos; particularmente RDF y OWL son los dos lenguajes dominantes de la Web Semántica. SPARQL (Angles & Gutierrez, 2008) es el lenguaje de consulta estándar por defecto para RDF.

- **SPARQL:** Es un lenguaje de consulta de OWL, debido a que OWL puede ser serializado como RDF. De manera que SPARQL puede ser usado para consultar el grafo OWL serializado en RDF. No obstante SPARQL no entiende de manera nativa OWL ya que este opera sólo en la serialización RDF y no tiene conocimiento de los constructores del lenguaje. Como resultado, SPARQL no puede consultar directamente las declaraciones hechas usando estos constructores. A pesar de esto hay una variedad de lenguajes derivados de SPARQL (Buil-Aranda et al., 2013), como RQL (Gregory Karvounarakis.), nRQL (Volker Haars lev, 2004), nSPARQL (Pérez, Arenas, & Gutierrez, 2008), SPARQL- DL (Sirin & Parsia, 2007), iSPARQL (Kiefer, Bernstein, Lee, Klein, & Stocker, 2007), SPARQL++ (Polleres, Scharffe, & Schindlauer, 2007), PSSPARQL y SPARQL-ST (Perry, Jain, & Sheth, 2011), SquishQL (Miller, Seaborne, & Reggiori, 2002). Los cuales cumplen con las necesidades básicas de consulta en aplicaciones muy específicas.

- **SQWRL**: Es un lenguaje de consulta nativo para OWL, está basado en el lenguaje de reglas SWRL (Semantic Web Rule Language) (Ian Horrocks, 2004). La principal diferencia entre SQWRL y SPARQL radica en que las consultas SPARQL se basan en la sintaxis de la tripleta sujeto-predicado-objeto, con lo cual no cubre el significado de la declaración, mientras que SQWRL trabaja directamente con las declaraciones en lugar de la sintaxis y abarca el modelo inferido. También ofrece una sintaxis muy cercana a la lógica de predicados, de manera que permite abstraer la estructura de la declaración subyacente.

2.1.14 Adopción de Buenas Prácticas

Para la producción de datos enlazados se han establecido directrices acerca de cómo estos deben ser publicados. Precisamente la W3C estableció un comité (W3C Working Group, 2006), el cual tiene como objetivo actualizar las recomendaciones relacionadas con la publicación de datos enlazados (Heath & Bizer, 2011).

Algunas de estas son:

- Abrir los datos a la web en cualquier formato pero bajo licencias libres.
- Hacer los datos accesibles como datos estructurados.
- Estructurar los datos bajo formatos no propietarios.
- Usar URIs para identificar lo publicado.
- Enlazar los enlazados con otros datos.

2.1.15 Ciclo de Vida de Datos Enlazados (Linked Data)

Este modelo brinda metodologías y herramientas para la publicación de datos en la web (LOD2, 2013).

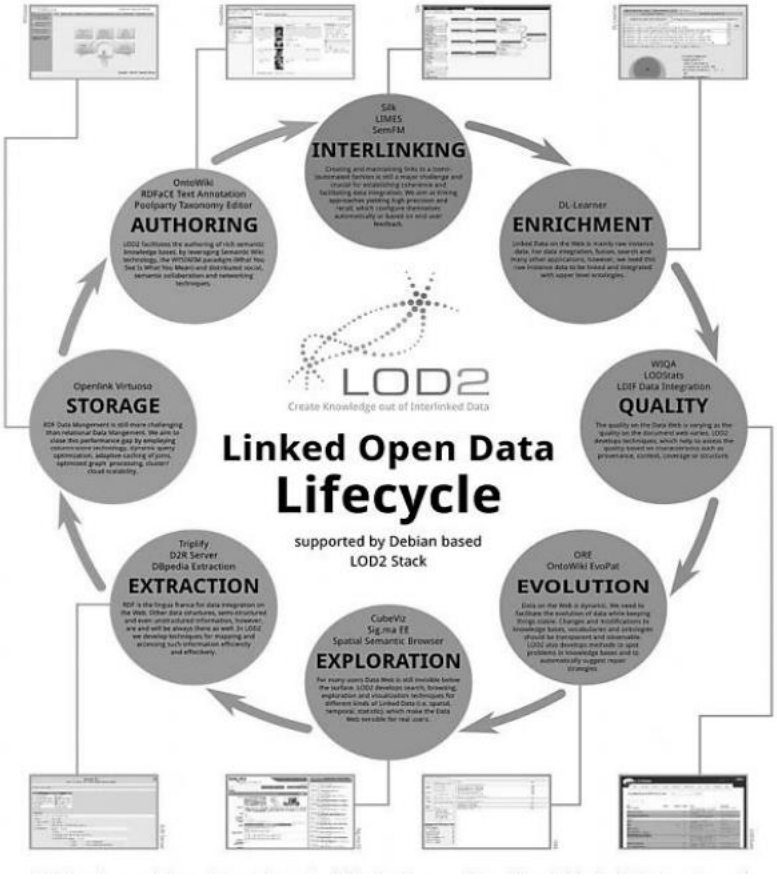


Figura 5. Linked Open Data Lifecycle

2.2 Marco Contextual

A continuación se indican algunas experiencias desarrolladas a nivel internacional sobre la extracción de información, modelos semánticos, almacenamientos en ontologías y web semántica de manera que este trabajo pueda ser contextualizado y examinado teniendo como referencia la situación actual.

Actualmente existen trabajos contenidos en la literatura que utilizan expresiones regulares (Regex) para la extracción de segmentos de texto desde fuentes semiestructuradas:

2.2.1 Modelado Semántico de Documentos con Estructura Definida

Osvaldo Salinas (Martínez Salinas, 2012), propone un método para la extracción de información a partir de documentos de texto semiestructurado mediante el uso de expresiones regulares para la representación semántica de la información en ontologías de dominio bajo los lineamientos de Linked Data.

El método consiste de cuatro tareas:

1) Extracción de información,
2) Poblado del modelo semántico,
3) Razonamiento e inferencia de información.
4) Persistencia del conocimiento.

2.2.2 Information Extraction from Web Pages

(Novotny, Vojtas, & Maruscak, 2009) se presentan una serie de técnicas para la extracción de datos de atributos de objetos de las páginas web que contengan datos o múltiples datos relacionados de uno o varios objetos. Esta información se extrae con la ayuda de una ontología de extracción.

2.2.3 Ontology Driven Discovery of Geospatial Evidence in Web Pages

Borges (Borges, Jr, Laender, & Medeiros, 2010), presenta un enfoque basado en ontologías que tiene como objetivo reconocer, extraer y codificar (parcial o completamente) evidencia geográfica en páginas web con como: calles, monumentos, códigos de área telefónica y códigos postales.

De manera que pueda proporcionar soporte a servicios basados en la localización, integrando páginas web con localizaciones urbanas, es decir páginas web concernientes a una localización urbana determinada mediante el análisis de su contenido. Combinando el enfoque de ontologías de extracción con diccionarios geográficos como GeoNames (Wick, 2012), obtienen una ontología denominada OnLocus.

Con la ontología resultante se ofrece la capacidad de reconocer y clasificar cadenas de caracteres de evidencia geográfica.

En esta propuesta se utilizan expresiones regulares para identificar direcciones a través de patrones sintácticos en notación EBNF ((ISO), 1996) tal como se muestra en la Figura 6.

Basic Address pattern

```
<ADDRESS>::= [<IDENT>] (<STREET_TYPE> <STREET_NAME> | ("BR" | "BR.") {0-9}+)
             [<S1>] <NUM>

<IDENT>::= ("Address:" | "ADDRESS:"| "ADDR:"| "Addr:"| "ADDR.:"| "Addr.:"|
           "Location:"| "Place:");

<STREET_TYPE>::= ("Street" | "STREET" | "St." | "St" |
                  "Avenue" | "AVENUE"| "Av" |"AV" | "Av."|"AV." |"Ave"|"AVE" |
                  "Highway" | "HIGHWAY" | "Hwy" | "HWY" | "Hwy." |"HWY." |
                  "Square" | "Sq"| "SQ"| "Sq."| "SQ."|
                  "Parkway" | "PARKWAY"| "Pkwy."| "PKWY."| "Pkwy"| "PKWY"|
                  "Road" | "ROAD" | "Rd"| "RD"| "Rd."| "RD."|
                  "Drive" | "DRIVE" | "Dr" | "DR" |"Dr." | "DR.");

<STREET_NAME>::= ({1-9} <PREP>  {(A| B |...|Z) {a-z}}+ |
                  {{(A| B |...|Z) {a-z}}+ <PREP>  {(A| B |...|Z) {a-z}}+}+ |
                  {(A| B |...|Z) {a-z}}+ {1-9}|
                  {{(A| B |...|Z)}+ {a-z}}+|
                  {1-9}+ ) ;

<PREP>::= ("of" | "on" |  "at"|  "in" | "of the" | "   ")  ;

<S1>::= (" , " |  " - " | "    ");

<NUM>::= ([<N>] {(0|1|...|9)}+ " . " (0|1|..|9) (0|1|..|9) (0|1|..|9) {a-z} |
         {A-Z})  |
         [<N>] {(0|1|...|9)}+ ({a-z} | {A-Z}) ) ;

<N>::= ("n." | "N." |"n" | "N" | "#" | "Mile Marker"| "Number" | "number");
```

Figura 6. Patrón utilizando notación EBNF ((ISO), 1996)

2.2.4 Método General de Extracción de Información Basado en el Uso de Lógica Borrosa. Aplicación en Portales Web

Jorge Ropero (Ropero Rodríguez, 2009), en su tesis propone el desarrollo de un agente inteligente capaz de responder a las necesidades de los usuarios en su proceso de encontrar la información deseada en entornos en los que la información es ingente, heterogénea, vaga, imprecisa o desordenada.

El principal aporte de esta tesis es la creación de un método general para la búsqueda y extracción de información basado en el uso de la Lógica Borrosa (Fuzzy Logic, FL), la cual es una herramienta ideal para la gestión de una información de las características antes mencionadas, y, en particular, la aplicación y validación de este método para la

extracción de información de portales web, dado que los portales web son un claro exponente de información heterogénea y desordenada.

2.2.5 Diseño de un Sistema de Extracción de Información de Artículos de Wikipedia

Miguel Sáez (Saez, 2009), propone el diseño de un sistema de extracción automática de información a partir de grandes corpus de textos. El desarrollo de este se centró en la búsqueda de información específica dentro de artículos de personajes contenidos en Wikipedia. El sistema diseñado tratará de establecer todas las relaciones posibles entre el artículo analizado y una serie de conceptos contenidos dentro del mismo (enlaces a otros artículos). Estas relaciones serán automáticamente clasificadas dentro de la categoría que se estime más adecuada (relación laboral, invención, lugar de residencia, etc.).

La implementación del sistema combina el uso de distintas técnicas de Procesamiento de Lenguaje Natural (incluyendo herramientas de análisis morfológico, sintáctico y semántico), la potencia de PHP para el procesamiento de textos de gran tamaño y la flexibilidad de las expresiones regulares tipo Perl.

2.2.6 iDocument: Using Ontologies for Extracting and Annotating Information from Unstructured Text

(Adrian, Hees, Van Elst, & Dengel, 2009), presentan una trabajo que pretende atacar el problema de la enorme cantidad de datos de texto en la WWW, Analizando formalmente la incorporación de ontologías de dominio en las tareas de extracción de información con iDocument, utilizando ontologías de dominio con el conocimiento del dominio formal y estructurado para la extracción de información desde un dominio de texto no anotada y no estructurado.

iDocument proporciona una arquitectura de canalización, una interfaz para el uso plantillas de extracción y capacidad de intercambiar ontologías de dominio para realizar tareas de extracción de información. También implementa el uso de consultas SPARQL.

2.2.7 Ontology-Based Information Extraction and Integration from Heterogeneous Data Sources

En este paper (Buitelaar, Cimiano, Frank, Hartung, & Racioppa, 2008) presentan el diseño, implementación y evaluación de SOBA, el cual es un sistema basado en Ontologias para la extracción de Informacion desde fuentes de datos heterogéneas, lo que incluye texto plano, tablas, imágenes y subtítulos.

También en el estado del arte de este trabajo se han podido identificar algunos trabajos relacionados con Web Semantica (Creación y Poblado de Ontologias) tales como:

2.2.8 Minerva: A Scalable Owl Ontology Storage and Inference System

En el artículo (Zhou et al., 2006), mediante cuatro módulos se realiza la importación, almacenamiento, inferencia y consulta de ontologías. El módulo de importación de ontologías traslada la ontología importada a una base de datos mediante el uso de Jena.

El módulo de almacenamiento se encarga de persistir tanto la información como la información inferida en la base de datos. El módulo de inferencia se encarga de obtener información no declarada explícitamente mediante el uso de un razonados DL y mediante reglas DB SQL. Por último, para consultar la ontología se usa SPARQL.

2.2.9 SABIOS: Una Aplicación de la Web Semántica Para la Gestión de Documentos Digitales

El artículo (Luna, A, Torres Pardo, & Ovalle, 2007), se muestra las bondades de la Web Semántica en la solución de los problemas que se presentan en la gestión de documentos en un entorno universitario.

La aplicación SABIOS plantea solución a la necesidad en escuela de Artes Plásticas no existía una terminología y un modelo de clasificación para los procesos de catalogación y uso de los documentos digitales que ellos generaban, de manera que se posibilitara un lenguaje unificado y facilitar así las labores de cada usuario en la publicación y recuperación de información dentro del sistema.

2.2.10 SWRC Ontology Semantic Web for Research Communities

SWRC (Sure, Bloehdorn, Haase, Hartmann, & Oberle, 2005), es una ontología para la descripción de comunidades de investigadores, ya que las comunidades semánticamente también representan organizaciones.

La ontología propuesta cubre las necesidades de representación semántica de la información relacionada a una organización pero no realiza el poblado a partir de una fuente textual sólo presentan los criterios de diseño de la ontología y ofrecen instrucciones para su reutilización.

2.2.11 Prototipo de un sistema de integración de recursos científicos, diseñado para su funcionamiento en el espacio de los datos abiertos enlazados para mejorar la colaboración, la eficiencia y promover la innovación en Colombia

En esta tesis (Quiroga, 2015), presenta la construcción de un prototipo así como la definición y desarrollo de un ciclo de vida para la producción de datos enlazados, mediante el uso de un conjunto de datos en el dominio de los recursos de investigación, apoyado en una amplia revisión de tecnologías y herramientas para su ejecución.

2.2.12 Development of Anti-Diabetic Drugs Ontology for Guideline Based Clinical Drugs Recommend System Using OWL and SWRL

En este trabajo (Chen, Bau, & Huang, 2010), se desarrolló una ontología para almacenar conocimientos de medicina, así como datos personales del paciente. Se utilizó lenguaje (SWRL) para construir las reglas relacionadas con los medicamentos anti-diabéticos.

Luego el conocimiento inmerso en esta ontología se transformó en un formato que permitiera el razonamiento de información relacionado con los medicamentos antidiabéticos adecuados, así como su monitorización, contraindicaciones y efectos secundarios.

2.2.13 Knoocks - A Visualization Approach for OWL Lite Ontologies

(Kriglstein & Wallner, 2010) presentan una herramienta para la visualización de ontologías OWL-Lite, esta herramienta está destinada para ontologías con una jerarquía muy extendida y un gran número de instancias.

(Ben Mustapha, Zghal, Aufaure, & ben Ghezala, 2010) en su trabajo, presentan un novedoso método para el alineamiento de ontologías del mismo contexto pero sintácticamente diferentes mediante el cálculo de similitudes locales y globales de la estructura del grafo.

2.2.14 CA Manager Framework: Creating Customized Workflows for Ontology Population and Semantic Annotation

(Damljanovic, Amardeilh, & Bontcheva, 2009) presenta un framework para crear distintos flujos de trabajo para la población de ontologías utilizando notación semántica basados en las recomendaciones de la Web Semántica y los preceptos UIMA (OASIS, 2009) .

Este framework soporta la población textual de ontologías de forma semiautomática; Además incluye herramientas y plugins de extracción de información, ontologías de dominios personalizados, y diversos repositorios semánticos, ofreciendo flexibilidad, sin comprometer la precisión y la recuperación de los componentes.

También en el estado del arte del proyecto se localizó literatura relacionada con consulta de Informacion para la web semantica siendo de gran utilidad para este trabajo:

2.2.15 Query for Semantic Web Services Using SPARQL-DL

(Wang, Zhai, & Fan, 2009) proponen utilizar semántica lenguaje de consulta SPARQL Web -DL como el lenguaje para la solicitud de recuperación de información de forma OWL- S desde servicios publicados.

De manera que permite utilizar consultas en SPARQL-DL para identificar servicios web definidos en OWL (OWL-S) (Burstein, 2004) o en semánticamente similares, así confirmando que este método es práctico, simple, fiable y fácil de usar.

3. Desarrollo e Implementación

En los capítulos anteriores se pretendió abarcar la introducción a los conceptos esenciales que permitiesen comprender la propuesta, mediante la presentación en el marco contextual de algunas experiencias afines encontradas en la literatura moderna.

En el capítulo 3 se presenta una propuesta para definir un método general que permita representar la información existente en documentos de texto semiestructurado al interior del Servicio Nacional de Aprendizaje SENA, en un modelo semántico siguiendo los lineamientos de Linked Data.

Para alcanzar satisfactoriamente este objetivo, se ha optado por seguir la metodología expuesta por Osvaldo Salinas, para el modelamiento semánticos de documentos con estructura definida, en la cual se presentan cuatro fases principales:

1. **Fase de Extracción de Información:** Esta actividad consiste en la extracción de información a partir de documentos de texto semiestructurado al interior del Servicio Nacional de Aprendizaje SENA, mediante el uso de patrones junto a expresiones regulares.

2. **Fase de Poblado del Modelo Semántico:** Esta actividad se realiza mediante la obtención de plantillas de extracción, las cuales están basadas en las propiedades de los elementos semánticos (clases de una ontología y sus propiedades). (debo habar de las limitantes de los documentos en PDF).

3. **Fase de Obtención de Información:** Esta actividad pretende representar el dominio utilizando las cualidades de razonamiento del lenguaje utilizado (RDF), mediante la aplicación de reglas de inferencia.

4. **Fase de Persistencia de conocimiento:** Esta actividad garantiza que la las instancias y la información inferida puedan ser persistentes y para esto se plantea el uso de algún estándar reconocido (RDF/XML), considerando las herramientas disponibles para administrar el modelo.

A grandes rasgos, estas actividades pretenden garantizar el poblado de un modelo semántico usando como insumo fundamental documentos de texto semiestructurado generados al interior del Servicio Nacional de Aprendizaje SENA, las cuales permitan realizar exitosamente la Fase de razonamiento e inferencia de información.

El flujo de trabajo de esta propuesta de acuerdo a la clasificación de los patrones de publicación de Linked Data se ha establecido como se observa en la Figura 7 de acuerdo con (Richard y Heath-Tom Bizer, 2007), el área bordeada en rojo delimita el alcance de esta propuesta.

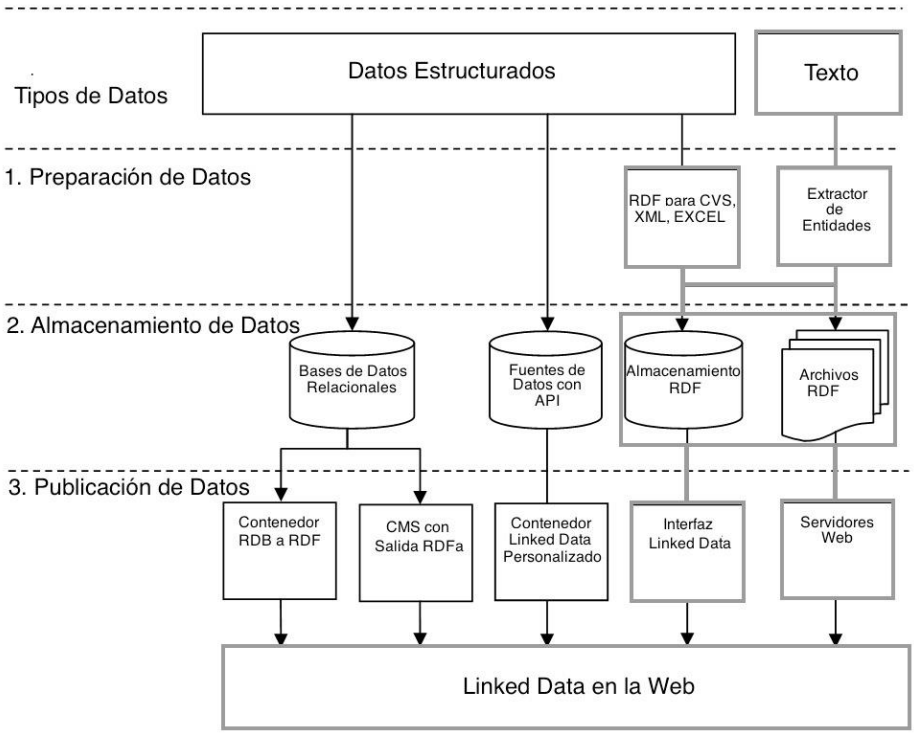


Figura 7. Alcance de la propuesta según el esquema de publicación de Linked Data

3.1 Caso de Prueba

En particular este caso de prueba se desarrolló considerando la necesidad de modelar, representar, extraer, inferir y consultar información al interior del Servicio Nacional de Aprendizaje SENA, donde existen gran cantidad de archivos, los cuales contienen información textual de manera semiestructurada, lo cual dificulta realizar consultas sobre la información allí contenida, impidiendo que pueda ser utilizada de manera activa al interior de la entidad.

Aunque actualmente la entidad posee un avanzado gestor documental, el cual se encarga de gestionar, almacenar e indexar los documentos producidos por procesos realizados al interior de la entidad, la información que se puede extraer de los mismos es bastante limitada, obligando en muchas ocasiones a abrir el documento para poder observar con mayor detalle el contenido en su interior.

En el SENA actualmente existen definidos tipos documentales en la plataforma OnBase (Hyland, 1991) de manera que en el futuro cualquiera de estos es una podría ser usado como fuente de información, ya que son documentos de texto semiestructurados.

Para este trabajo se seleccionó el tipo de documental CONTRATOS, y el Subtipo Contratos Servicios (Prestación de Servicios personales), ya que actualmente el SENA cuenta con más de 28000 contratistas a nivel nacional, y en muchos casos un mismo contratista puede tener durante un año tres o más contratos con la entidad lo que aumenta esta cifra drásticamente y esto se debe multiplicar por la cantidad de años hacia atrás desde que se cuenta con el gestor documental.

	40		CONTRATOS	X	X		1	19	X	X		X			Cumplido el tiempo de retención se hace una selección cualitativa, de aquellos contratos cuya cuantía y objeto incidieron en el cumplimiento de la misión de la entidad y generan valor secundario como fuente histórica de consulta
		01	Contratos de Adquisición de Bienes y Servicios (Compraventa, Suministro, Mantenimiento, Obra: Licitación, Convocatoria Pública e Invitación Pública, selección Abreviada, Concurso de Meritos, Contratación Directa y Mínima Cuantía)												
			Etapa Preparatoria												

CONVENCIONES						
TRADICIÓN DOCUMENTAL	DISPOSICIÓN FINAL				SÍMBOLOS	
Soporte Físico: ORIGINAL O COPIA	CT= CONSERVACIÓN TOTAL	SF= Soporte Físico	SE= Soporte Electrónico	S=SELECCIÓN	SERIE= MAYÚSCULA FIJA NEGRILLA	Subserie = Mayúscula Inicial negrilla
Soporte Electrónico: BASE DE DATOS	Tecnología de Conservación	D= Digitalización	M= MICROFILMACIÓN	E= ELIMINACIÓN	• = Tipo Documental	

SECRETARIA GENERAL

Grupo de Administración de Documentos 3

SECRETARÍA GENERAL – Grupo Administración de Documentos *Versión 2013*

Vo. Bo. Nohra Patricia Vela Otálora, Coordinadora Grupo Administración de Documentos - *Aprobada mediante Resolución No. 0577 del 27/03/2014*

Figura 8. Tipo documental contratos en el SENA.

Además la indexación de estos documentos, en la mayoría de los casos se realiza 100% manual, lo que expone a la entidad a errores humanos debidos a los altos volúmenes de documentos generados, así como a las múltiples fuentes que los generan, lo que impide que la información histórica contenida en estos documentos sea utilizada eficazmente como soporte en la toma de decisiones de la entidad.

Actualmente la plataforma de gestión documental OnBase permite indexar los documentos mediante un formulario el cual contiene campos relacionados con la información propia del contrato (Keywords).

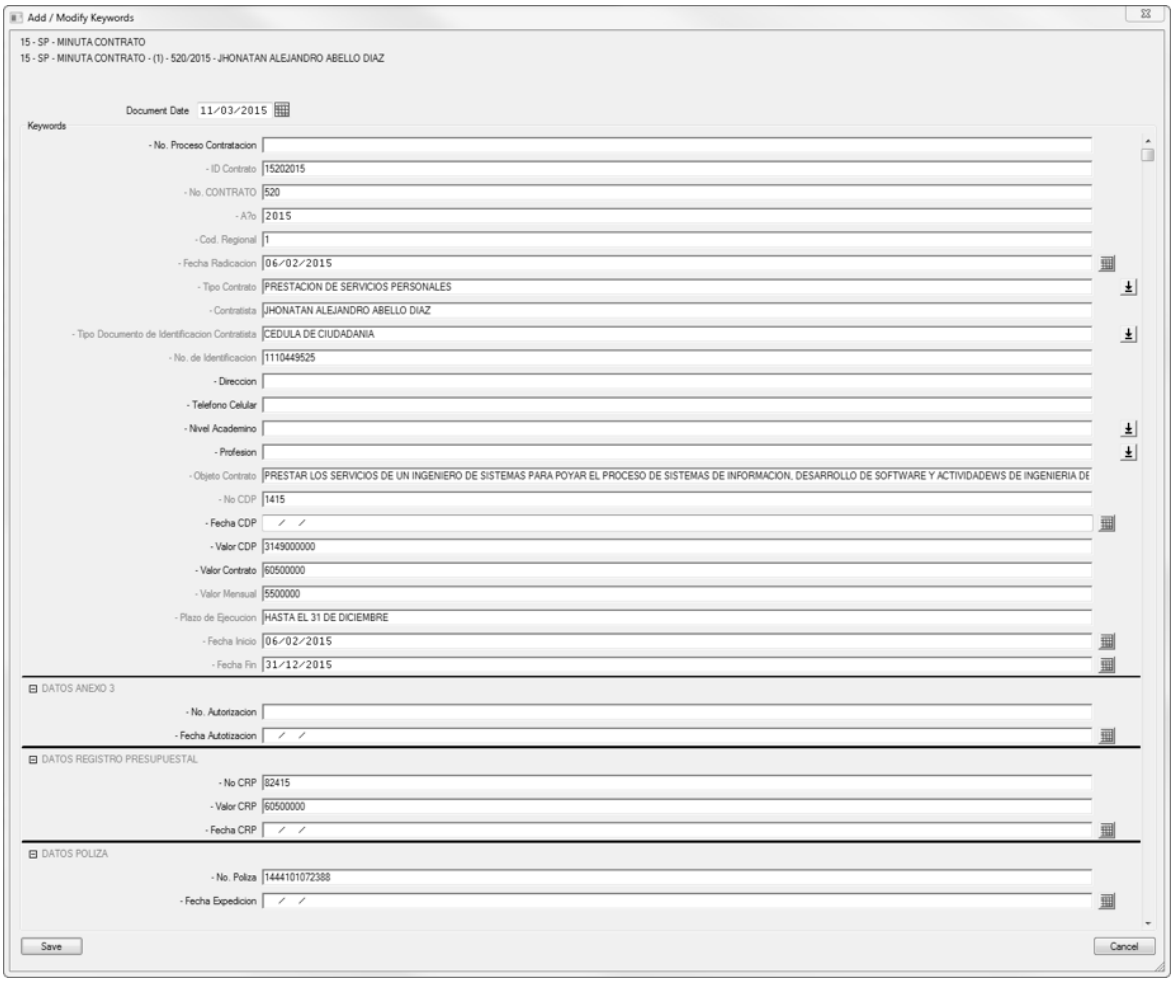


Figura 9. Formulario Add/Modify Keywords OnBase

Para tener una idea concreta de la estructura de estos documentos en el Anexo A se muestra de cómo está organizada la información de acuerdo a unos lineamientos predefinidos

3.2 Metodología

La metodología propuesta en este trabajo final de maestría pretende ofrecer unas pautas que faciliten desarrollar una aplicación para la Web Semántica siguiendo los principios fundamentales de Linked Data, utilizando como insumo básico información textual, la cual pueda ser transformada en conocimiento que pueda ser procesado y consumido por diferentes sistemas de información.

Para lograr esto es necesario modelar el dominio de conocimiento, es decir representarlo en un modelo semántico, introducir la información extraída de los documentos de texto semiestructurados, razonar e inferir información a partir del modelo semántico y definir un mecanismo de consulta (SPARQL) (Buil-Aranda et al., 2013).

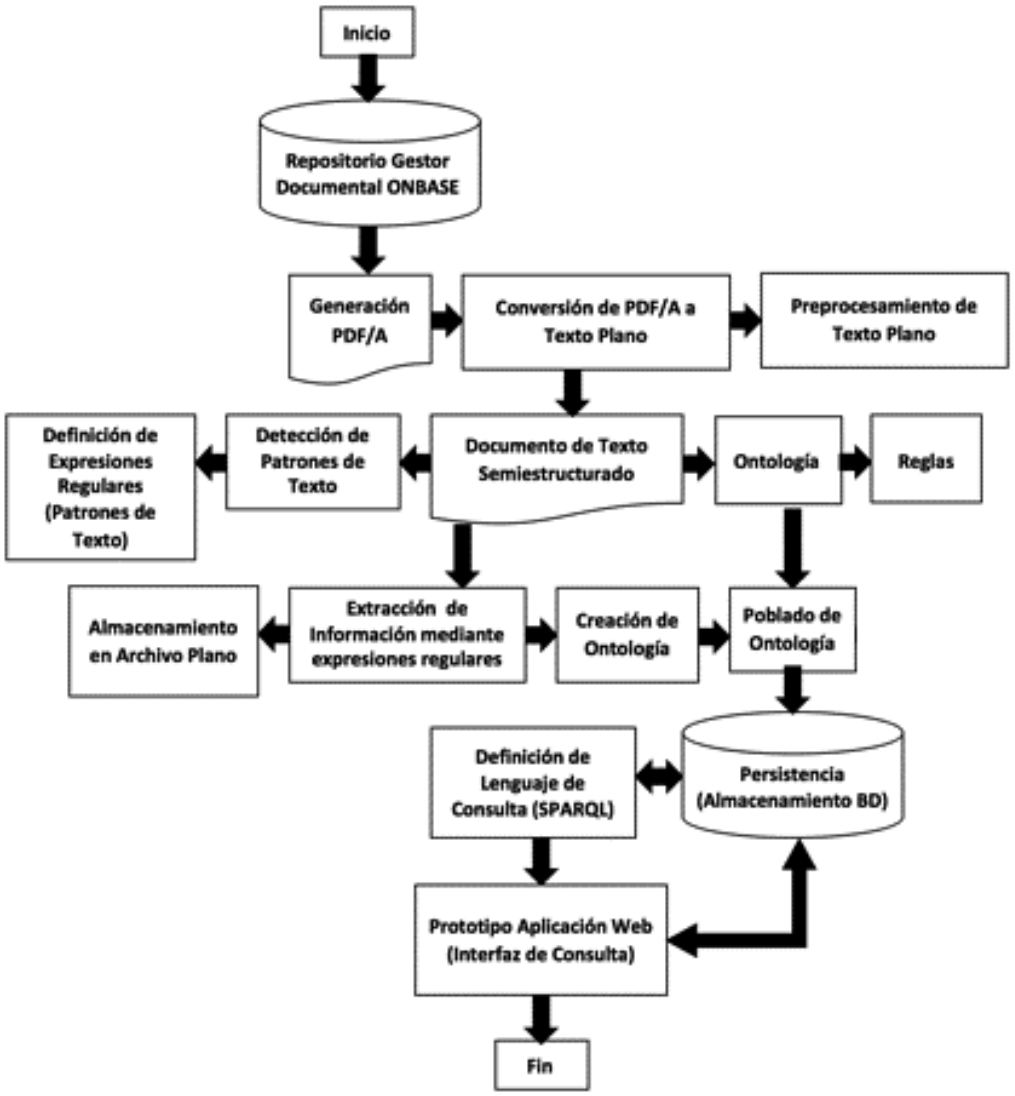


Figura 10. Diagrama de Flujo del método propuesto para extracción de información, poblado de ontologías y consulta de información basado en los principios de Linked Data.

3.2.1 Entorno de Desarrollo del Prototipo de Solución

Las especificaciones de hardware y software de la maquina utilizada son:

Figura 11. Especificaciones de hardware y software

De esta manera se pretende desarrollar un prototipo de aplicación que permita la extracción de información a partir de documentos semiestructurados producidos al interior del Servicio Nacional de Aprendizaje SENA, permitiendo su publicación, reutilización e intercambio siguiendo los lineamientos y estándares de la web semántica publicados por la W3C.

3.3 Flujo de Procesos Extracción de Información

A continuación se detalla el flujo de procesos necesarios para realizar la extracción de información:

3.3.1 Obtención de Documentos del Repositorio ONBASE

Actualmente el SENA implemento como gestor documental la herramienta OnBase (Hyland, 1991). Tal como se describe en su página oficial, esta “permite a las organizaciones automatizar procesos de negocios, reducir el tiempo y el costo de

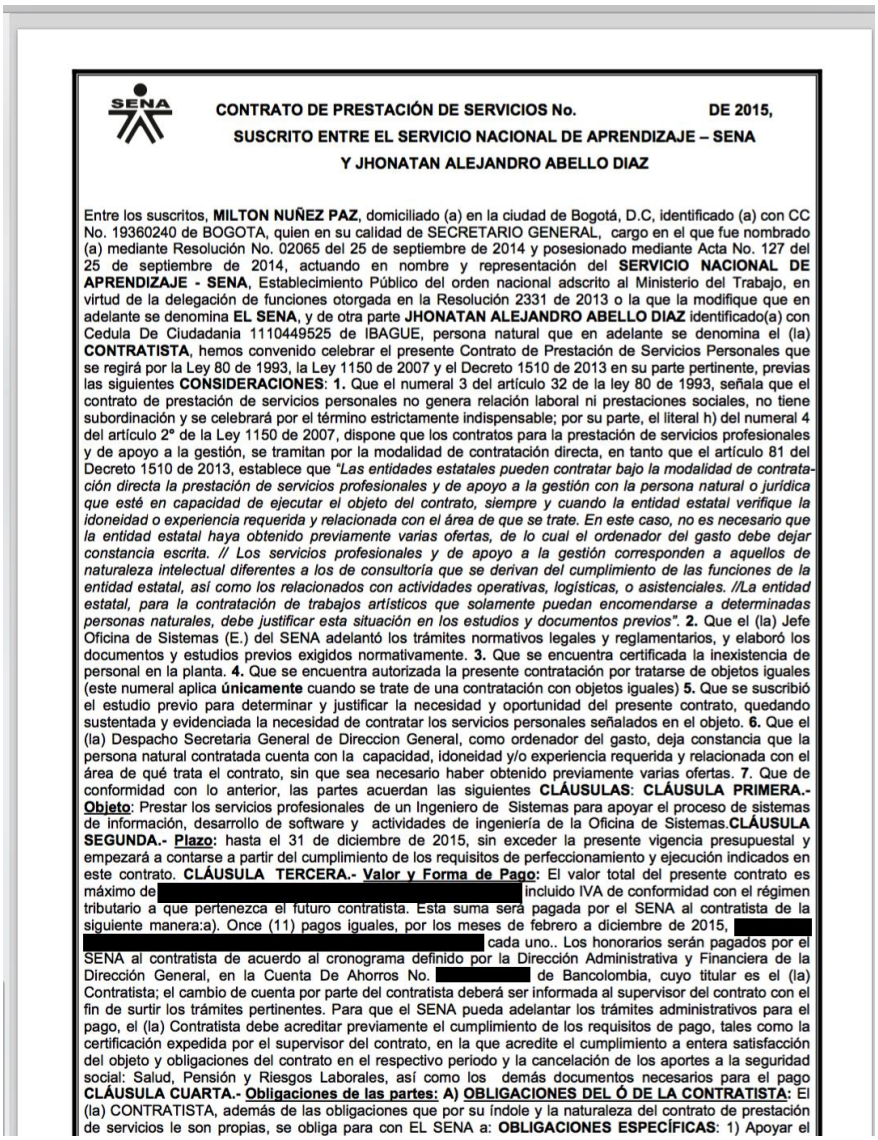
efectuar funciones de negocios importantes y mejorar la eficiencia organizacional. También es una herramienta estratégica para cumplir con normas ISO, SOX o gubernamentales por medio del manejo y control de contenido desde cualquier origen. Una de las grandes ventajas de OnBase es que facilita el intercambio de documentos e información en una interface intuitiva y simple con empleados, socios de negocio y clientes".

3.3.2 Generación de PDF/A

Es necesario generar la información desde OnBase en el formato (PDF/A Competence Center, 2011), el cual es un formato de archivo para el archivo a largo plazo de documentos electrónicos. Este brinda la posibilidad de que la información en su interior pueda ser extraída, lo cual es de suma utilidad para el desarrollo de este trabajo.

3.3.3 Conversión de PDF/A a Texto Plano

Esta tarea consiste en extraer las cadenas de caracteres que representan el texto desde el archivo fuente, que para nuestro caso concreto es PDF.



CONTRATO DE PRESTACIÓN DE SERVICIOS No. DE 2015,
SUSCRITO ENTRE EL SERVICIO NACIONAL DE APRENDIZAJE – SENA
Y JHONATAN ALEJANDRO ABELLO DIAZ

Entre los suscritos, **MILTON NUÑEZ PAZ**, domiciliado (a) en la ciudad de Bogotá, D.C, identificado (a) con CC No. 19360240 de BOGOTA, quien en su calidad de SECRETARIO GENERAL, cargo en el que fue nombrado (a) mediante Resolución No. 02065 del 25 de septiembre de 2014 y posesionado mediante Acta No. 127 del 25 de septiembre de 2014, actuando en nombre y representación del **SERVICIO NACIONAL DE APRENDIZAJE - SENA**, Establecimiento Público del orden nacional adscrito al Ministerio del Trabajo, en virtud de la delegación de funciones otorgada en la Resolución 2331 de 2013 o la que la modifique que en adelante se denomina **EL SENA**, y de otra parte **JHONATAN ALEJANDRO ABELLO DIAZ** identificado(a) con Cedula De Ciudadania 1110449525 de IBAGUE, persona natural que en adelante se denomina el (la) **CONTRATISTA**, hemos convenido celebrar el presente Contrato de Prestación de Servicios Personales que se regirá por la Ley 80 de 1993, la Ley 1150 de 2007 y el Decreto 1510 de 2013 en su parte pertinente, previas las siguientes **CONSIDERACIONES**: **1.** Que el numeral 3 del artículo 32 de la ley 80 de 1993, señala que el contrato de prestación de servicios personales no genera relación laboral ni prestaciones sociales, no tiene subordinación y se celebrará por el término estrictamente indispensable; por su parte, el literal h) del numeral 4 del artículo 2° de la Ley 1150 de 2007, dispone que los contratos para la prestación de servicios profesionales y de apoyo a la gestión, se tramitan por la modalidad de contratación directa, en tanto que el artículo 81 del Decreto 1510 de 2013, establece que *"Las entidades estatales pueden contratar bajo la modalidad de contratación directa la prestación de servicios profesionales y de apoyo a la gestión con la persona natural o jurídica que esté en capacidad de ejecutar el objeto del contrato, siempre y cuando la entidad estatal verifique la idoneidad o experiencia requerida y relacionada con el área de que se trate. En este caso, no es necesario que la entidad estatal haya obtenido previamente varias ofertas, de lo cual el ordenador del gasto debe dejar constancia escrita. // Los servicios profesionales y de apoyo a la gestión corresponden a aquellos de naturaleza intelectual diferentes a los de consultoría que se derivan del cumplimiento de las funciones de la entidad estatal, así como los relacionados con actividades operativas, logísticas, o asistenciales. //La entidad estatal, para la contratación de trabajos artísticos que solamente puedan encomendarse a determinadas personas naturales, debe justificar esta situación en los estudios y documentos previos".* **2.** Que el (la) Jefe Oficina de Sistemas (E.) del SENA adelantó los trámites normativos legales y reglamentarios, y elaboró los documentos y estudios previos exigidos normativamente. **3.** Que se encuentra certificada la inexistencia de personal en la planta. **4.** Que se encuentra autorizada la presente contratación por tratarse de objetos iguales (este numeral aplica **únicamente** cuando se trate de una contratación con objetos iguales) **5.** Que se suscribió el estudio previo para determinar y justificar la necesidad y oportunidad del presente contrato, quedando sustentada y evidenciada la necesidad de contratar los servicios personales señalados en el objeto. **6.** Que el (la) Despacho Secretaria General de Direccion General, como ordenador del gasto, deja constancia que la persona natural contratada cuenta con la capacidad, idoneidad y/o experiencia requerida y relacionada con el área de qué trata el contrato, sin que sea necesario haber obtenido previamente varias ofertas. **7**. Que de conformidad con lo anterior, las partes acuerdan las siguientes **CLÁUSULAS**: **CLÁUSULA PRIMERA.- Objeto**: Prestar los servicios profesionales de un Ingeniero de Sistemas para apoyar el proceso de sistemas de información, desarrollo de software y actividades de ingeniería de la Oficina de Sistemas.**CLÁUSULA SEGUNDA.- Plazo:** hasta el 31 de diciembre de 2015, sin exceder la presente vigencia presupuestal y empezará a contarse a partir del cumplimiento de los requisitos de perfeccionamiento y ejecución indicados en este contrato. **CLÁUSULA TERCERA.- Valor y Forma de Pago:** El valor total del presente contrato es máximo de [redacted] incluido IVA de conformidad con el régimen tributario a que pertenezca el futuro contratista. Esta suma será pagada por el SENA al contratista de la siguiente manera:a). Once (11) pagos iguales, por los meses de febrero a diciembre de 2015, [redacted] cada uno.. Los honorarios serán pagados por el SENA al contratista de acuerdo al cronograma definido por la Dirección Administrativa y Financiera de la Dirección General, en la Cuenta De Ahorros No. [redacted] de Bancolombia, cuyo titular es el (la) Contratista; el cambio de cuenta por parte del contratista deberá ser informada al supervisor del contrato con el fin de surtir los trámites pertinentes. Para que el SENA pueda adelantar los trámites administrativos para el pago, el (la) Contratista debe acreditar previamente el cumplimiento de los requisitos de pago, tales como la certificación expedida por el supervisor del contrato, en la que acredite el cumplimiento a entera satisfacción del objeto y obligaciones del contrato en el respectivo periodo y la cancelación de los aportes a la seguridad social: Salud, Pensión y Riesgos Laborales, así como los demás documentos necesarios para el pago **CLÁUSULA CUARTA.- Obligaciones de las partes: A) OBLIGACIONES DEL Ó DE LA CONTRATISTA:** El (la) CONTRATISTA, además de las obligaciones que por su índole y la naturaleza del contrato de prestación de servicios le son propias, se obliga para con EL SENA a: **OBLIGACIONES ESPECÍFICAS**: 1) Apoyar el

Figura 12. Contrato PDF/A

Para realizar la tarea de extracción de texto plano existen una gran variedad de APIs, algunas Open Source y otras desarrolladas con propósitos comerciales. En este trabajo se utilizara PDF Parser (sebastien malot, 2013), la cual es una Liberia desarrollada en PHP que permite extraer elementos de texto desde un archivo PDF.

La documentación de la misma puede ser localizada en la web del autor http://www.pdfparser.org/documentation.

3.3.4 Preprocesamiento de Texto Plano

Esta tarea consiste en detectar caracteres extraños y/o patrones para convertirlos a su correspondiente equivalente en la codificación compatible con el motor de búsqueda.

Además también es conveniente eliminar saltos de línea, los cuales comúnmente son generados debido a longitud de los párrafos y saltos de página, ya que en diferentes sistemas es representado de distintas maneras “\n” (Linux), “\r\n” o puede no ser considerarlo como patrón de texto.

Debido a que el texto extraído se encuentra en español, el cual tiene algunas particularidades propias tales como: caracteres adicionales y acentuación, al realizar la extracción se debe garantizar que los caracteres sean codificados en un formato compatible con Regex. Para nuestro caso en particular se ha seleccionado UTF-8, debido a que:

- Permite representar cualquier carácter Unicode.
- Usa símbolos de longitud variable (de 1 a 4 bytes por carácter Unicode).
- Incluye la especificación US-ASCII de 7 bits, permitiendo que cualquier mensaje ASCII sea representado sin cambios.
- Incluye sincronía, lo cual permite determinar el inicio de cada símbolo sin reiniciar la lectura desde el principio.
- No superposición.

También es de suma importancia la organización del documento (estructura) en la que se presentan las cadenas de caracteres, ya que en lo posible deben seguir una secuencia comprensible. Por ejemplo, si la extracción de información está basada en el patrón etiqueta-valor, es ideal que la cadena etiqueta esta contigua a la cadena valor, ya que de lo contrario disminuirá la precisión de la extracción.

3.3.5 Detección de Patrones de Texto

Tal como se mencionó en la tarea de Preprocesamiento de texto plano, éste debe contener párrafos, frases o estructuras de texto coherentes. Comúnmente el texto plano extraído de documentos de texto semiestructurado puede contener alguno de estos dos tipos de patrones:

3.3.5.1 Patrones de Texto Etiqueta - Valor

Este tipo de patrón permite identificar de manera explícita información, por ejemplo los datos personales de una persona, como nombre, apellidos, edad, etc. Comúnmente este tipo de patrón está constituido por una cadena de texto con la etiqueta y otra con el valor; Tradicionalmente separados por signos de puntuación tales como: “:”, ”,”, ”-”.

A continuación se presenta una muestra de datos del contrato generados, los cuales presentan patrones del tipo etiqueta - valor:

Etiqueta	Valor
Objeto	Prestar los servicios profesionales de un
Contrato de	Prestacion de Servicios
Cedula de Ciudadania	1110449525
Disponibilidad Presupuestal No.	225
Expedido	marzo 23 de 2015
Plazo	31 de Diciembre

Figura 13. Patrones del tipo etiqueta-valor de los datos relacionados con el contrato.

Una estrategia que permite identificar los patrones, es mediante el conteo de palabras que pueda contener o identificar la etiqueta contigua o el carácter de terminación. De esta manera se puede diseñar las reglas adecuadas para extraer la información exacta.

3.3.5.2 Patrones de Texto Estructural

Existen otro tipo de patrones de texto, los cuales son comúnmente encontrados en forma de lista. Un caso usual es una lista de referencias bibliográficas, en la que cada entrada es similar a esta:

Decker S., Brickley D., Saarela, and Angele J., "A query and inference service for RDF", in QL'98 – The Query Languages Workshop, Boston, USA, 1998.

Para este caso si conocemos con anticipación la estructura de la anterior referencia bibliográfica es posible identificar que la referencia es de un paper y está constituida por autores, título, conferencia, lugar y año de publicación.

Para nuestro caso particular el desarrollo de este trabajo, no fue posible encontrar patrones diferentes a Patrones de texto etiqueta-valor

En este caso de prueba se identificaron 15 patrones diferentes para igual número de productos, estos productos se clasificaron en diversas categorías de acuerdo a criterios particulares de la Entidad.

De manera que un patrón puede identificar a los datos de más de un tipo de producto.

Por lo tanto es necesario para cada campo definir un patrón y un método de extracción.

```
if(preg_match($patrón, @$contenido))
{
    //echo "existeee";
    ?>
    <script language="javascript">
    var contenido = "<?php echo $contenido; ?>" ;
    var Pos1=contenido.indexOf(cdpContenido);
    var Pos2=contenido.indexOf(", expedido");
    var fechaCdpContenido=contenido.slice(Pos1+14,Pos2);
    alert(fechaCdpContenido);
    </script>
    <?php
}
else
    {
        echo "NO existeee";
    }
////////////////////////////////////////////////////////////
```

Figura 14. Extracción del CDP mediante patrones de texto etiqueta-valor y funciones en JavaScript.

Es el anterior caso del CDP, se implementaron expresiones regulares para verificar la existencia del patrón de búsqueda al interior del documento, después se procede a utilizar las funciones JavaScript indexOf, lastIndexOf y Slice, las cuales nos permiten realizar el trim del contenido de nuestro interés y almacenarlo en una variable para su posterior procesamiento.

Es importante al realizar esta actividad, considerar las posibles variantes en longitud y contenido de cada valor, así como las posibilidades de términos sinónimos o abreviaturas en la etiqueta. También es importante utilizar el modificador /i, el cual no diferencian mayúsculas y minúsculas aumentando las posibilidades de acertar en nuestra búsqueda. Otra buena práctica es intercambiar las letras acentuadas por su similar sin acento o tildes. Esto se debe a que los operadores no los incluyen como parte de la clase alfanumérica ya que son considerados caracteres especiales.

3.3.6 Documento de Texto Semiestructurado

Se realiza la carga del contenido de texto en un archivo plano luego del Preprocesamiento de texto plano y se procede a realizar la búsqueda de información mediante el uso de expresiones regulares y patrones.

El conjunto de documentos usado en el caso de prueba consta de 22 archivos en formato PDF, este conjunto de documentos fue generado mediante el gestor documental OnBase. Como ya se mencionó, estos documentos contienen información referente a los datos personales de los contratistas, así como los datos propios de los contratos de prestación de servicios.

Es importante aclarar que estos documentos al contener información sensible fue necesario en algunas imágenes y adjuntos anonimizar mediante una franja negra cierta información confidencial.

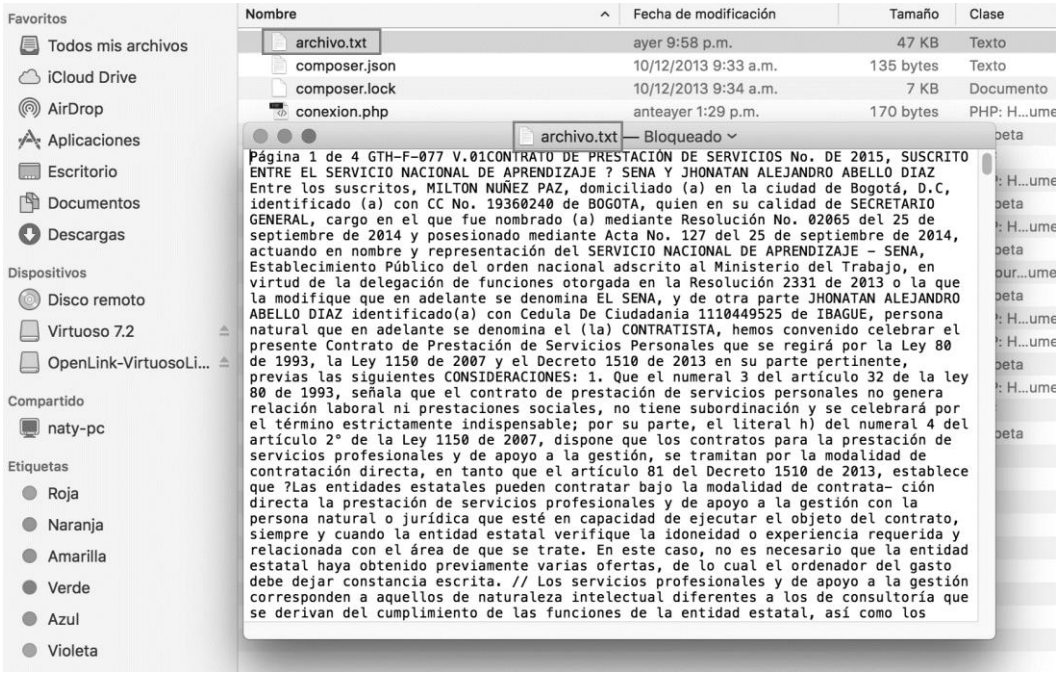


Figura 15. Contenido almacenado en .txt del documento de Texto Semiestructurado

3.3.7 Modelado Semántico y Poblado de Ontología

El poblado de una ontología consiste en agregar instancias al modelo. Esta tarea requiere definir previamente el modelo semántico, de manera que es necesario modelar el dominio de conocimiento para representarlo en alguno de los lenguajes existentes para este fin (World Wide Web Consortium, 2012) y luego poblar el modelo semántico con la información obtenida con anterioridad.

3.3.7.1 Modelo de Datos Semántico

Con el objetivo de representar adecuadamente el dominio de conocimiento, existen varios modelos y lenguajes formales para la definición de la semántica de la información y así convertirla en conocimiento, entre estos encontramos vocabularios, taxonomías, tesauros, mapas de tópicos, etc. Para el desarrollo de este trabajo se emplearan ontologías como modelo semántico, el cual permite representar el dominio de conocimiento.

Ontología, es un modelo para la representación de conocimiento formal basado en redes semánticas. Este modelo de datos está compuesto por clases, propiedades, relaciones, restricciones y axiomas, estas características permiten que una ontología sea más formal que otros modelos de datos (vocabularios, taxonomías, tesauros, mapas de tópicos, etc.).

El lenguaje para la definición de ontologías web OWL (World Wide Web Consortium, 2012) es ampliamente usado en la Web Semántica. Gracias a esto han surgido una serie de buenas prácticas para la construcción de ontologías OWL, como las publicadas (Noy & McGuinness, 2012), en donde mediante una metodología simple de 7 pasos, se indican los elementos básicos a tener en cuenta en el modelado ontológico del dominio de conocimiento. Además actualmente podemos contar con herramientas de software que ayudan a construir, editar y visualizar ontologías OWL de manera manual o asistida; entre estas destacan Open Refine (“OpenRefine”, 2012), Protégé (“The Protege Project”, 2000), Sesame (Broekstra et al., 2002; “Sesame”, 2012), Jena (Carroll et al., 2004), TopBraid Composer.

3.3.7.2 Ontología de Dominio

Linked Data recomienda la reutilización de ontologías o estructuras semánticas, ya que mediante éstas es posible establecer un vocabulario común entre las aplicaciones que las utilizan.

Entre las estructuras semánticas más reutilizadas para la representación semántica de información sobre dominios específicos como relaciones interpersonales, material bibliográfico y comunidades de investigadores se encuentran:

- **FOAF** (Friend of a Friend, 2012): Acrónimo de Friendo of a Friendo, define un diccionario de propiedades y clases nombradas en RDF. Este vocabulario es usado para definir información personal y sus relaciones sociales. Este precisamente es el vocabulario utilizado en el desarrollo de este trabajo como veremos más adelante.

- **SWRC** (Sure et al., 2005): Es una ontología para representar comunidades de investigadores tales como personas, organizaciones, publicaciones y sus relaciones.

- **Dublín Core** (Dublin Core, 2012): Es un conjunto de clases y propiedades para descripción de material bibliotecario. Este vocabulario es usado para representar metadatos de material literario como revistas, publicaciones, libros, etc.

Siguiendo la metodología propuesta en este trabajo es necesario contar con una ontología que represente el conocimiento del dominio, para esto tenemos dos caminos a elegir, construirla desde cero o reutilizar una ontología ya existente garantizando que esta ajuste a las necesidades particulares del trabajo y permita satisfacer las necesidades particulares de representación del conocimiento en el caso de prueba.

Actualmente las ontologías son utilizadas ampliamente en la Web Semántica para representar el significado de los recursos existentes en la Web. Para hacer posible la administración de los recursos representados en la ontología es necesario contar con herramientas que faciliten esta tarea. Para nuestro caso particular utilizaremos Protégé,

el cual es un entorno de desarrollo para la creación y edición de ontologías, además de ser un marco de trabajo para bases de conocimiento.

Protégé soporta el modelado de ontologías en OWL, además permite exportar el modelo a diversos formatos como: RDF, OWL, XML.

3.3.7.3 Ontología del Caso de Prueba

De acuerdo con la propiedad de extensibilidad, la reutilización es una actividad altamente recomendable en el diseño y construcción ontologías. En este sentido, la ontología FOAF es reutilizada en este trabajo, esto permitió diseñar una ontología propia que cubre las necesidades particulares del caso de prueba. A continuación se muestra la taxonomía de conceptos que a su vez representan las clases de recursos en el dominio de conocimiento a operar.

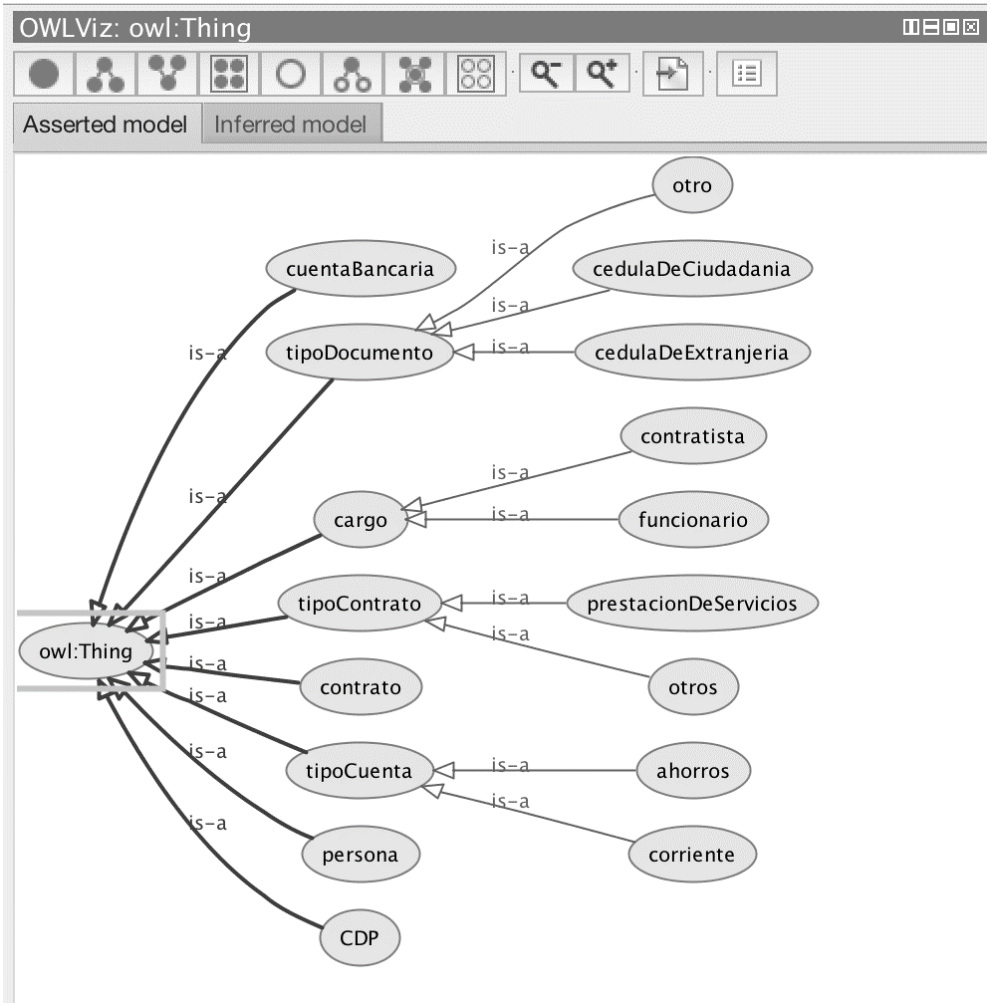


Figura 16. Vista parcial de la jerarquía de clase del caso de prueba.

Y luego podemos visualizar la reutilización de la ontología FOAF en el diseño de nuestro modelo.

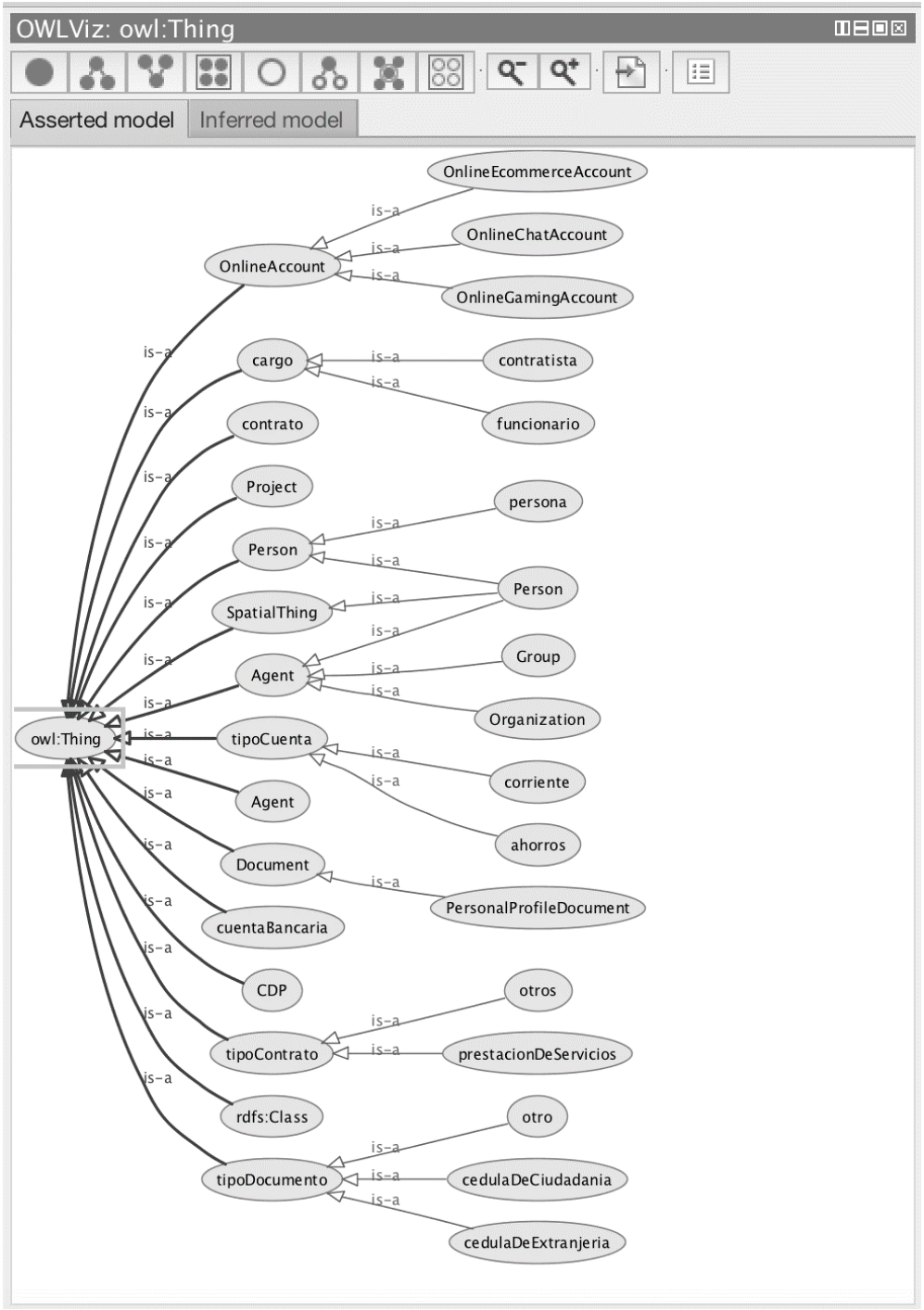


Figura 17. Vista la jerarquía de clase del caso de prueba reutilizando la ontología FOAF

Esta ontología se compone de 34 clases/conceptos, de los cuales 14 son principales (8 propias del caso de prueba) y 17 subclases (11 propias del caso de prueba), las cuales

están relacionadas de forma jerárquica, es decir, en los nodos cercanos a la raíz se encuentran los conceptos con mayor generalidad y a partir de éstos se establecen los conceptos más específicos pero que se encuentran dentro del ámbito del concepto padre.

Además de los conceptos se declararon 38 propiedades de objetos (ObjectProperties), así como diferentes propiedades de tipos de datos (DatatypeProperties), restricciones de rango y dominio, así como restricciones de cardinalidad.

3.3.7.4 Poblado de la Ontología

El poblado de la ontología se realiza con base en los datos extraídos y almacenados temporalmente.

En esta etapa se hace uso de la herramienta (“OpenRefine”, 2012), la cual permite cargar archivos .csv con la información extraída desde los documentos de texto semiestructurado, la cual fue almacenada previamente en la base de datos mydb creada en MySQL.

Figura 18. Vista de tablas Base de datos mydb

Y puede ser exportada fácilmente a .csv desde la opción importar de phpMyAdmin.

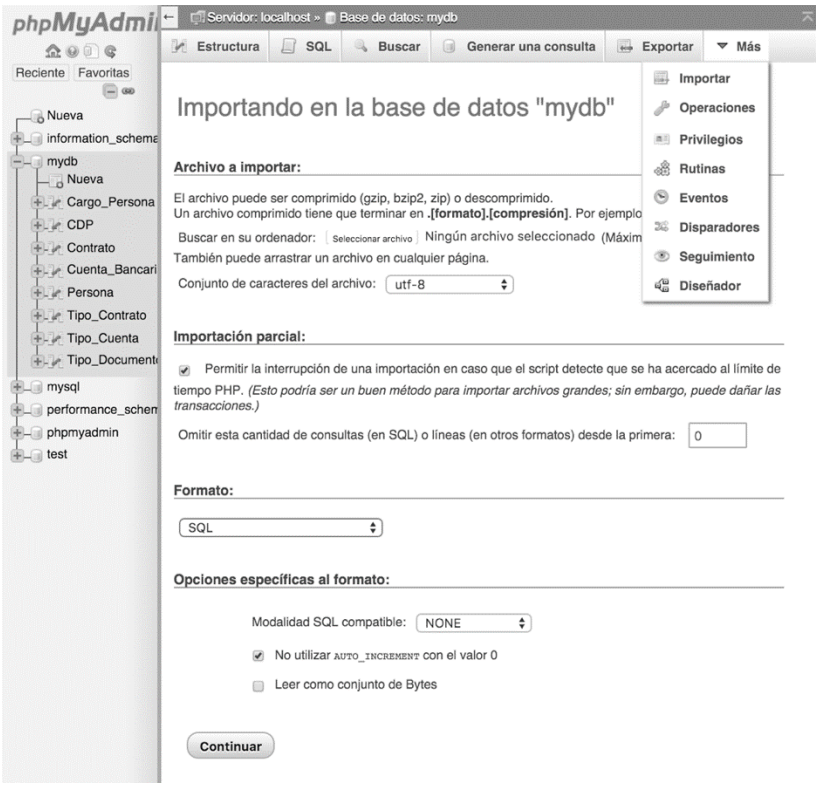


Figura 19. Importar contenido a csv desde phpMyAdmin

De esta manera se recuperan los diferentes elementos contenidos en la base de datos y posteriormente traducirlos a instancias de clases correspondiente en el grafo que representa la ontología.

3.3.8 Recuperación de Información

Esta tarea tiene como objetivo realizar una serie de razonamientos tanto en el modelo de datos como en las instancias las cuales permitan obtener información. Esta se realiza teniendo como base las características del modelo de datos de referencia, de manera que permita establecer que la información obtenida corresponde con la información alojada.

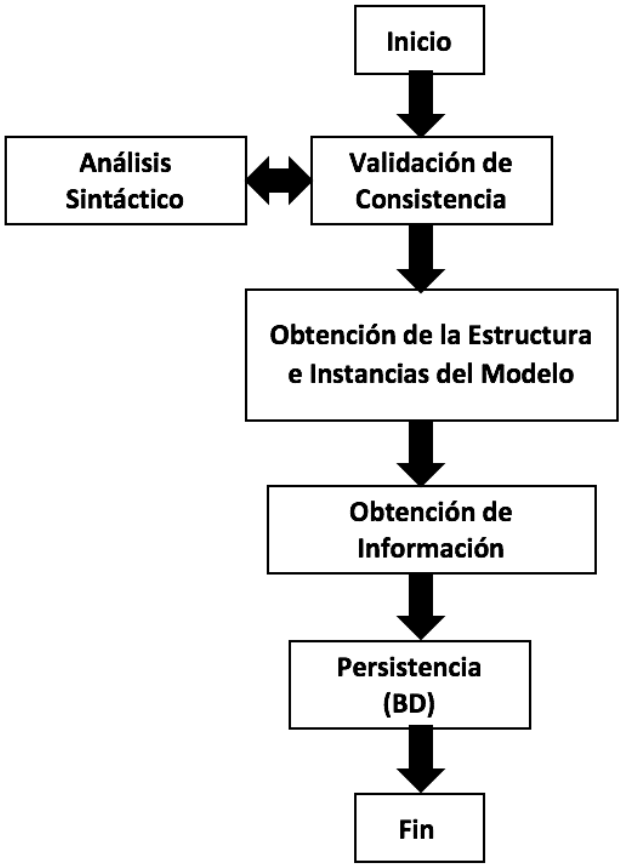
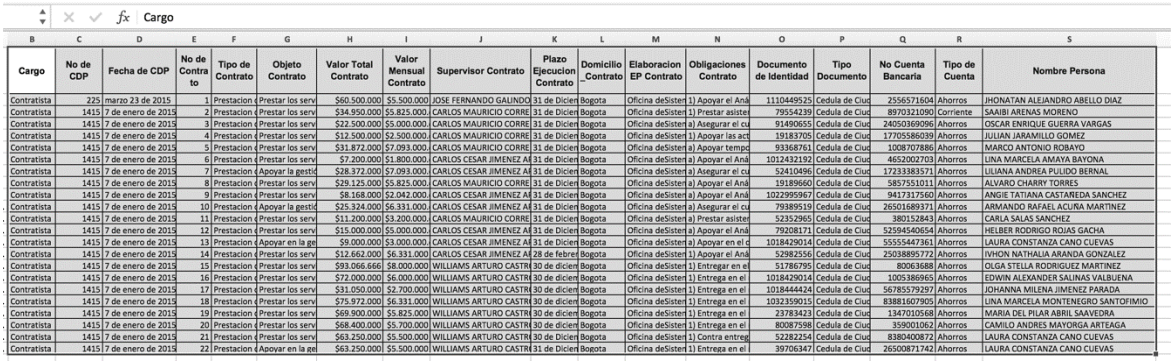


Figura 20. Diagrama de flujo razonamiento e inferencia de información

Una forma muy recurrida para organizar la información extraída es mediante plantillas de extracción. Estas plantillas son utilizadas como vehículo para trasladar los datos extraídos hacia el modelo de representación final, que en este caso en particular es una ontología de dominio. Una plantilla de extracción es una estructura tipo tabla donde se establece una serie de etiquetas para cada elemento extraído, para nuestro caso puntual se generó la siguiente plantilla de extracción:



Cargo	No de CDP	Fecha de CDP	No de Contrato	Tipo de Contrato	Objeto Contrato	Valor Total Contrato	Valor Mensual Contrato	Supervisor Contrato	Plazo Ejecucion Contrato	Domicilio Contrato	Elaboracion EP Contrato	Obligaciones Contrato	Documento de Identidad	Tipo Documento	No Cuenta Bancaria	Tipo de Cuenta	Nombre Persona
Contratista	225	marzo 23 de 2015	1	Prestacion (	Prestar los serv	$60.500.000	$5.500.000	JOSE FERNANDO GALINDO	31 de Dicien	Bogota	Oficina deSisten	1) Apoyar el Aná	1110449525	Cedula de Ciud	2556571604	Ahorros	JHONATAN ALEJANDRO ABELLO DIAZ
Contratista	1415	7 de enero de 2015	2	Prestacion (	Prestar los serv	$34.950.000	$5.825.000	CARLOS MAURICIO CORRE	31 de Dicien	Bogota	Oficina deSisten	1) Prestar asiste	79554239	Cedula de Ciud	8970321090	Corriente	SAAIBI ARENAS MORENO
Contratista	1415	7 de enero de 2015	3	Prestacion (	Prestar los serv	$22.500.000	$5.000.000	CARLOS MAURICIO CORRE	31 de Dicien	Bogota	Oficina deSisten	a) Asegurar el cu	91490655	Cedula de Ciud	24050369096	Ahorros	OSCAR ENRIQUE GUERRA VARGAS
Contratista	1415	7 de enero de 2015	4	Prestacion (	Prestar los serv	$12.500.000	$2.500.000	CARLOS MAURICIO CORRE	31 de Dicien	Bogota	Oficina deSisten	1) Apoyar las act	19183705	Cedula de Ciud	17705586039	Ahorros	JULIAN JARAMILLO GOMEZ
Contratista	1415	7 de enero de 2015	5	Prestacion (	Prestar los serv	$31.872.000	$7.093.000	CARLOS MAURICIO CORRE	31 de Dicien	Bogota	Oficina deSisten	a) Apoyar tempo	93368761	Cedula de Ciud	1008707886	Ahorros	MARCO ANTONIO ROBAYO
Contratista	1415	7 de enero de 2015	6	Prestacion (	Prestar los serv	$7.200.000	$1.800.000	CARLOS CESAR JIMENEZ AI	31 de Dicien	Bogota	Oficina deSisten	a) Apoyar el Aná	1012432192	Cedula de Ciud	4652002703	Ahorros	LINA MARCELA AMAYA BAYONA
Contratista	1415	7 de enero de 2015	7	Prestacion (	Apoyar la gestio	$28.372.000	$7.093.000	CARLOS CESAR JIMENEZ AI	31 de Dicien	Bogota	Oficina deSisten	a) Asegurar el cu	52410496	Cedula de Ciud	17233383571	Ahorros	LILIANA ANDREA PULIDO BERNAL
Contratista	1415	7 de enero de 2015	8	Prestacion (	Prestar los serv	$29.125.000	$5.825.000	CARLOS MAURICIO CORRE	31 de Dicien	Bogota	Oficina deSisten	a) Apoyar el Aná	19189660	Cedula de Ciud	5857551011	Ahorros	ALVARO CHARRY TORRES
Contratista	1415	7 de enero de 2015	9	Prestacion (	Prestar los serv	$8.168.000	$2.042.000	CARLOS CESAR JIMENEZ AI	31 de Dicien	Bogota	Oficina deSisten	a) Apoyar el Aná	1022995967	Cedula de Ciud	9417317560	Ahorros	ANGIE TATIANA CASTAÑEDA SANCHEZ
Contratista	1415	7 de enero de 2015	10	Prestacion (	Apoyar la gestio	$25.324.000	$6.331.000	CARLOS CESAR JIMENEZ AI	31 de Dicien	Bogota	Oficina deSisten	a) Asegurar el cu	79389519	Cedula de Ciud	26501689371	Ahorros	ARMANDO RAFAEL ACUÑA MARTINEZ
Contratista	1415	7 de enero de 2015	11	Prestacion (	Prestar los serv	$11.200.000	$3.200.000	CARLOS MAURICIO CORRE	31 de Dicien	Bogota	Oficina deSisten	a) Prestar asiste	52352965	Cedula de Ciud	380152843	Ahorros	CARLA SALAS SANCHEZ
Contratista	1415	7 de enero de 2015	12	Prestacion (	Prestar los serv	$15.000.000	$5.000.000	CARLOS CESAR JIMENEZ AI	31 de Dicien	Bogota	Oficina deSisten	a) Apoyar el Aná	79208171	Cedula de Ciud	52594540654	Ahorros	HELBER RODRIGO ROJAS GACHA
Contratista	1415	7 de enero de 2015	13	Prestacion (	Apoyar en la ge	$9.000.000	$3.000.000	CARLOS CESAR JIMENEZ AI	31 de Dicien	Bogota	Oficina deSisten	a) Apoyar en el c	1018429014	Cedula de Ciud	55555447361	Ahorros	LAURA CONSTANZA CANO CUEVAS
Contratista	1415	7 de enero de 2015	14	Prestacion (	Prestar los serv	$12.662.000	$6.331.000	CARLOS CESAR JIMENEZ AI	28 de febre	Bogota	Oficina deSisten	1) Apoyar el Aná	52982556	Cedula de Ciud	25038895772	Ahorros	IVHON NATHALIA ARANDA GONZALEZ
Contratista	1415	7 de enero de 2015	15	Prestacion (	Prestar los serv	$93.066.666	$8.000.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Entregar en el	51786795	Cedula de Ciud	80063688	Ahorros	OLGA STELLA RODRIGUEZ MARTINEZ
Contratista	1415	7 de enero de 2015	16	Prestacion (	Prestar los serv	$72.000.000	$6.000.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Entrega en el	1018429014	Cedula de Ciud	1005386965	Ahorros	EDWIN ALEXANDER SALINAS VALBUENA
Contratista	1415	7 de enero de 2015	17	Prestacion (	Prestar los serv	$31.050.000	$2.700.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Entrega en el	1018444424	Cedula de Ciud	56785579297	Ahorros	JOHANNA MILENA JIMENEZ PARADA
Contratista	1415	7 de enero de 2015	18	Prestacion (	Prestar los serv	$75.972.000	$6.331.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Entrega en el	1032359015	Cedula de Ciud	83881607905	Ahorros	LINA MARCELA MONTENEGRO SANTOFIMIO
Contratista	1415	7 de enero de 2015	19	Prestacion (	Prestar los serv	$69.900.000	$5.825.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Entrega en el	23783423	Cedula de Ciud	1347010568	Ahorros	MARIA DEL PILAR ABRIL SAAVEDRA
Contratista	1415	7 de enero de 2015	20	Prestacion (	Prestar los serv	$68.400.000	$5.700.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Entrega en el	80087598	Cedula de Ciud	359001062	Ahorros	CAMILO ANDRES MAYORGA ARTEAGA
Contratista	1415	7 de enero de 2015	21	Prestacion (	Prestar los serv	$63.250.000	$5.500.000	WILLIAMS ARTURO CASTR	30 de dicien	Bogota	Oficina deSisten	1) Contra entreg	52282254	Cedula de Ciud	8380400872	Ahorros	LAURA CONSTANZA CANO CUEVAS
Contratista	1415	7 de enero de 2015	22	Prestacion (	Apoyar en la ge	$63.250.000	$5.500.000	WILLIAMS ARTURO CASTR	31 de Dicien	Bogota	Oficina deSisten	1) Entrega en el	39706347	Cedula de Ciud	26500871742	Ahorros	LAURA CONSTANZA CANO CUEVAS

Figura 21. Plantilla general de extracción del caso de prueba

La validación de una ontología se realiza mediante la validación de consistencia y la inferencia u obtención de información de manera que estas actividades permitan validar que el diseño y la construcción de la ontología se realizaron correctamente.

Las principales características que se busca al realizar estas actividades de validación son:

- **Consistencia:** Asegura que una ontología no contenga hechos contradictorios.
- **Satisfactibilidad:** Determina si una clase puede tener instancias.
- **Clasificación:** Calcula la relación subclase entre todas las clases nombradas para crear la jerarquía de clase completa. Esta jerarquía puede ser usada para responder consultas tales como obtener todas o sólo las subclases directas de una clase.
- **Realización:** Es útil para encontrar la clase más específica a la que pertenece un individuo.

3.3.9 Consideraciones Previas a la Persistencia del Modelo

La persistencia del modelo semántico es de gran importancia, ya que mediante ésta es posible materializar el modelo y la información inferida evitando recalcular esta última en cada consulta, es decir trasladar la información obtenida a memoria secundaria permite

posteriormente realizar operaciones sobre la base de conocimiento resultante.

Para persistir la base de conocimiento existen varios formatos, tales como Turtle, N3, RDF/XML, entre otros.

Para nuestro caso particular del caso de prueba se usa el formato RDF/XML ya que es el formato por defecto utilizado por Open Refine y Virtuoso (OpenLink Software, 2012).

Antes de generar la persistencia de nuestro caso de prueba, es necesario recordar que:

- El modelo de persistencia seleccionado influye directamente en el desempeño de la aplicación final en donde se emplea la ontología. Para el desarrollo de este trabajo, es necesario contar con acceso a los datos mediante un lenguaje de consulta, que para nuestro caso puntual es SPARQL. Además también es necesario contar con la infraestructura de software que permita esto.

- También debemos tener en cuenta que en Linked Data el acceso a la información mediante URI es un hecho que se debe considerar en la persistencia de información.

En otras palabras la tarea de persistencia de conocimiento no sólo trata de materializar la base de conocimiento en disco sino que también trata de satisfacer los medios por los cuales se accederá a la información.

3.3.10 Persistencia del Modelo

3.3.10.1 Carga de Plantilla de Extracción en Open Refine

Mediante Open Refine podemos cargar directamente nuestra plantilla de extracción y crear un proyecto con esta información:

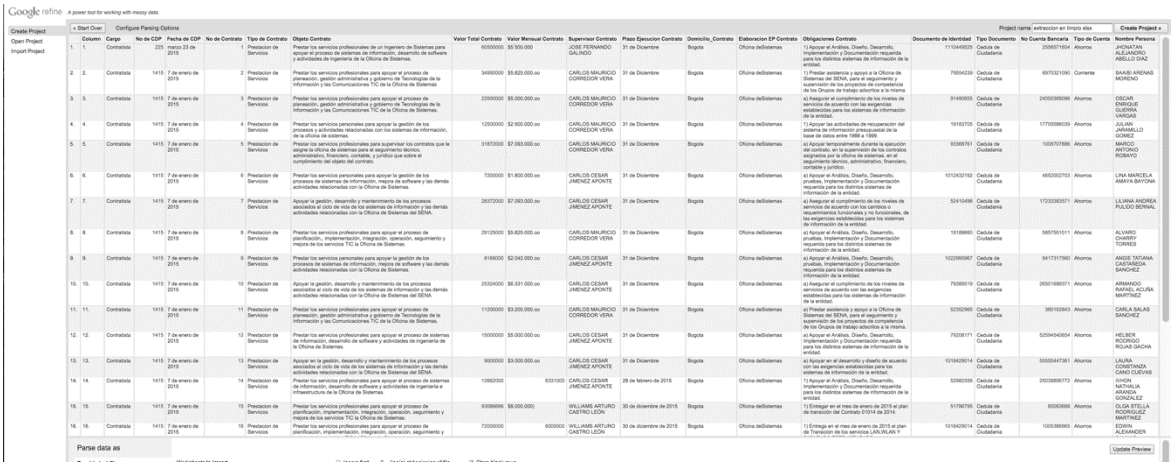


Figura 22. Plantilla de extracción en Open Refina

3.3.10.2 Interfaz de Extracción en Open Refine

A continuación podemos visualizar la interfaz de Open Refine que permite trabajar con la información obtenida.

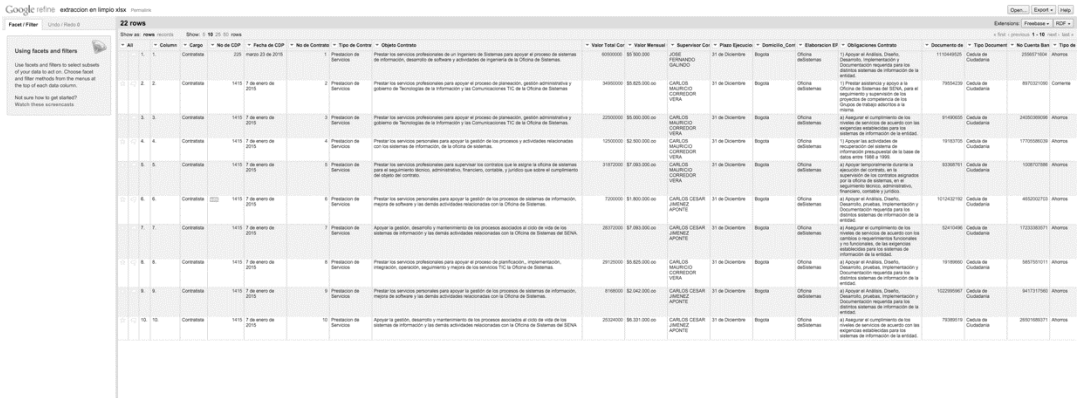


Figura 23. Creación de proyecto en Open Refine

3.3.10.3 Alinear Estructura del Esquema RDF en Open Refine

Luego debemos alinear la estructura (esqueleto) del esquema RDF, de manera que se especifique cómo los datos RDF serán generados a partir de los datos cargados y también debemos definir la Base URI, la cual para este trabajo es **http://localhost/prototipoTesis/**

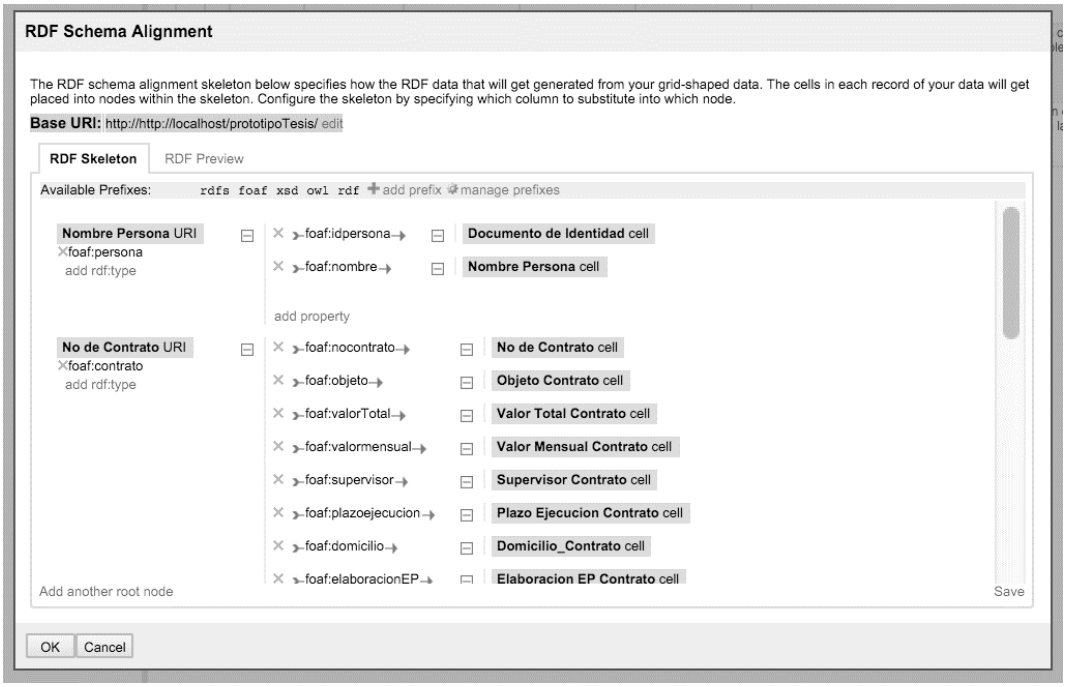


Figura 24. Alineación de estructura del esquema RDF en Open Refine

3.3.10.4 Servicio de Reconciliación en Open Refine

Luego procedemos a añadir el servicio de la reconciliación basada en archivos:

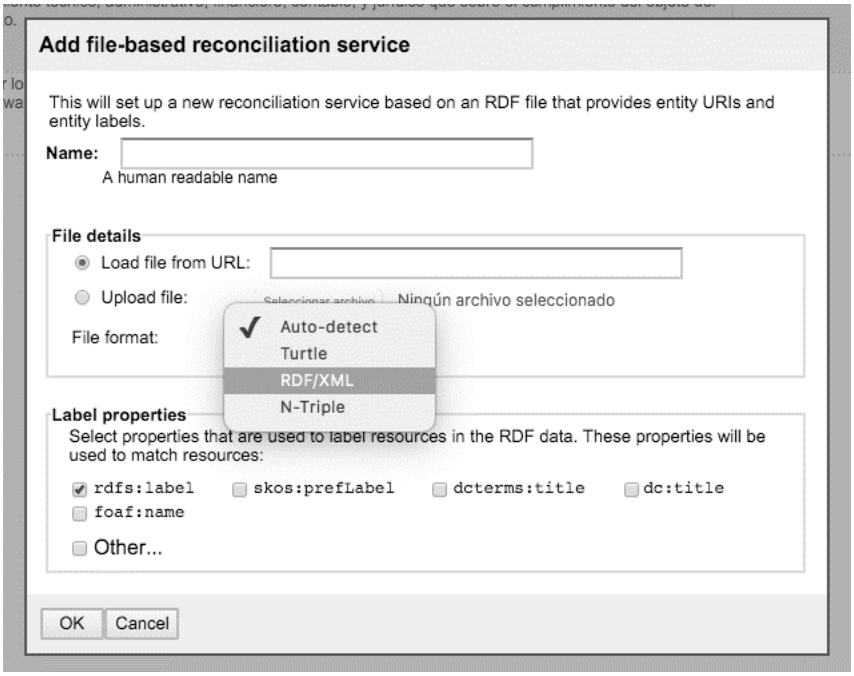


Figura 25. Añadir proceso de reconciliación en Open Refine

3.3.10.5 Proceso de Verificación de Ontología en Protege

Lo cual instanciar la información de acuerdo a la estructura definida en la ontología, lo cual se verifica al cargar nuestra ontología en PROTEGE y verificar que existen instancias de las clases definidas:

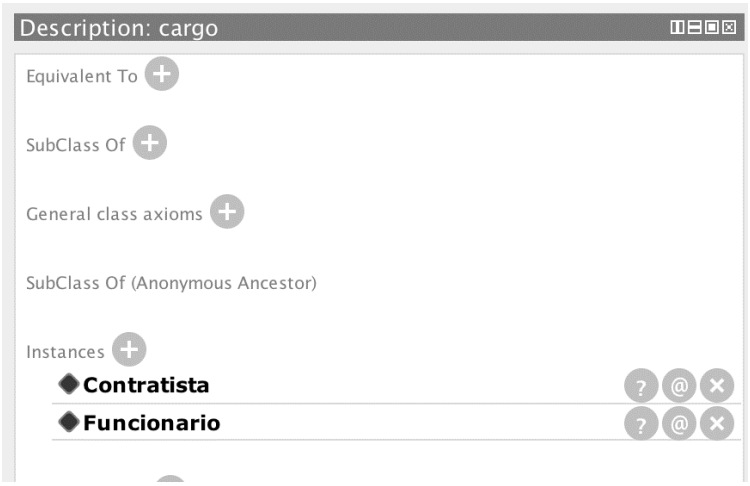


Figura 26. Verificación de la ontología del proyecto en Protege

3.3.10.6 Ejecución del Proceso de Reconciliación en Open Refine

Utilizando los servicios de reconciliación disponibles en Open Refine, se realizan los procesos de reconciliación para cada uno de los datos que conforman el caso de prueba.

Esto se hace mediante el cuadro de reconciliación dando clic en Start reconciling:

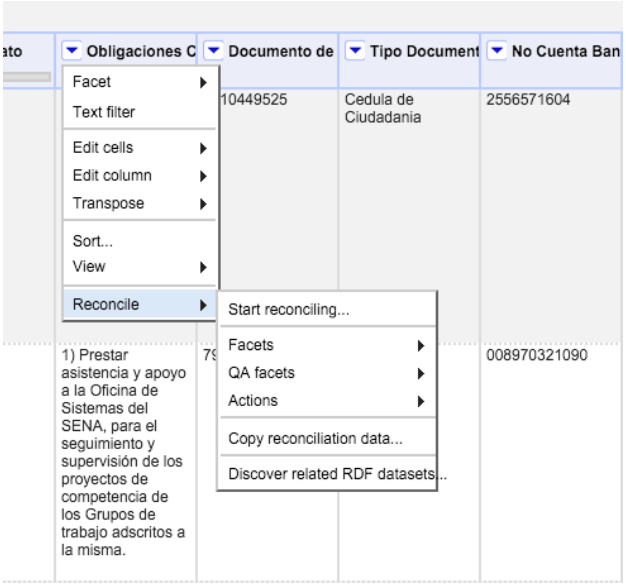


Figura 27. Configuración del proceso de reconciliación en Open Refine

Lo que permite que esta aplicación prueba con algunos de los valores contenidos en la columna y verifique si la información disponible en el servicio de reconciliación seleccionado puede determinar su tipo.

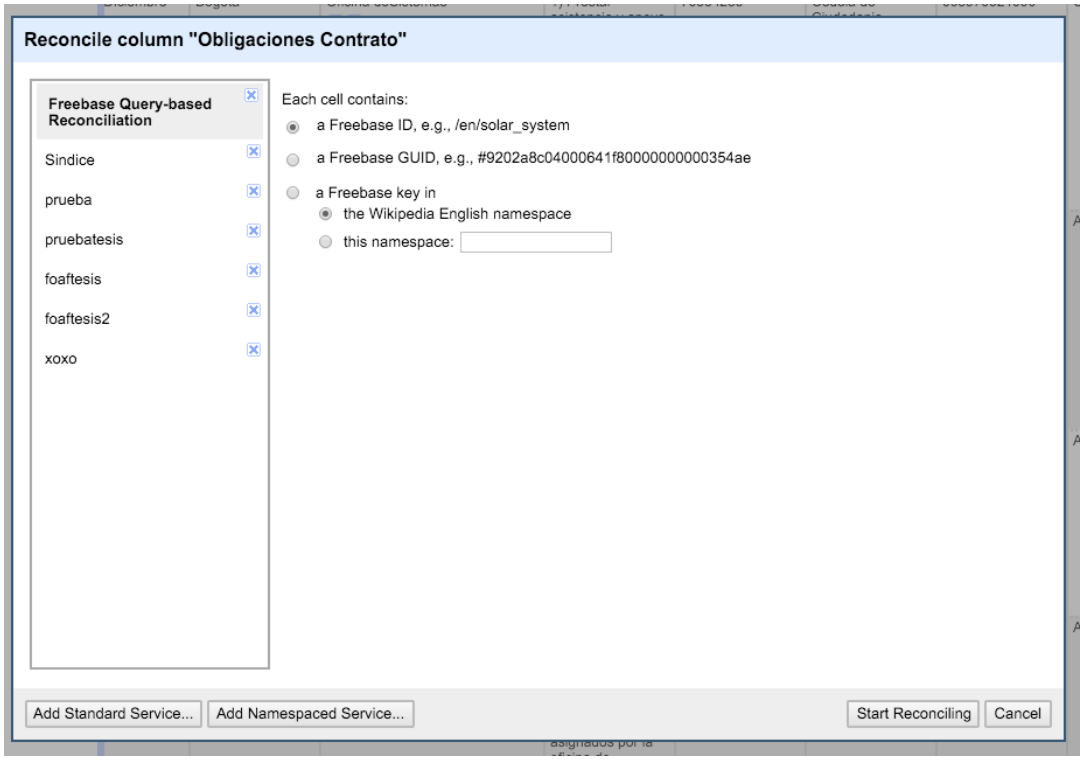


Figura 28. Reconciliación columna Obligaciones Contrato en Open Refine

Cuando el proceso de reconciliación finaliza, Open Refine añade una barra en la parte superior de la columna, que indica cuantas de las celdas han sido reconciliadas.

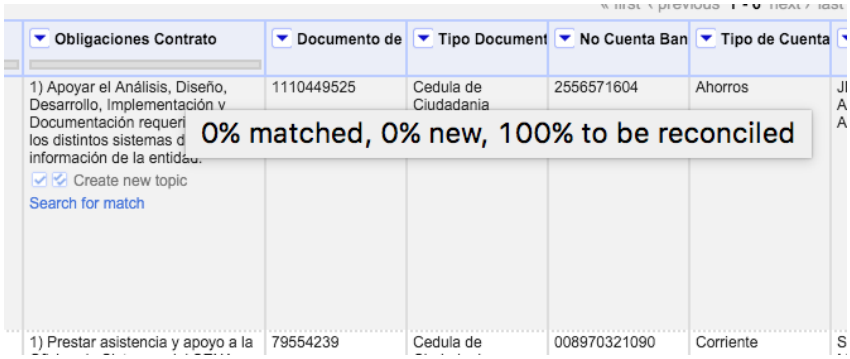


Figura 29. Barra de verificación en celdas reconciliadas en Open Refine

Desafortunadamente el tipo de información expuesta a través de Linked Data en nuestro trabajo no es de uso común, ya que se busca que el SENA sea pionero en Colombia en cuanto la exposición de datos contractuales en el sector público.

Posteriormente para generar mediante el proceso de reconciliación, la URL del conjunto de datos enlazados procedemos a añadir una columna derivada sobre la columna reconciliada.

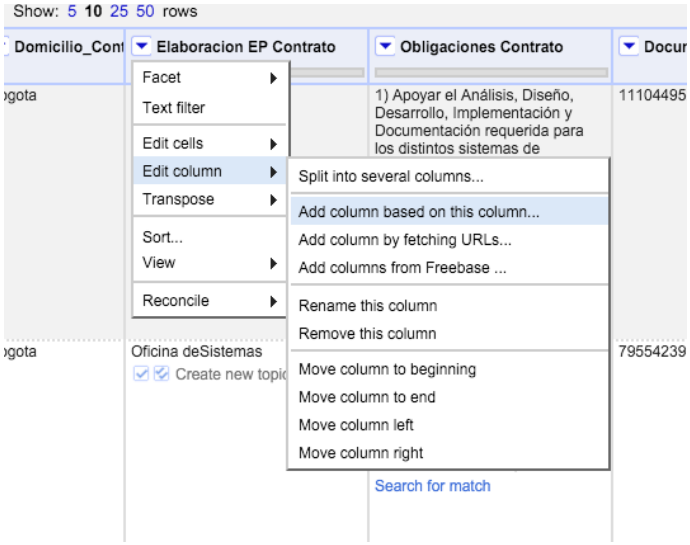


Figura 30. Selección de columna derivada en Open Refine

3.3.10.7 Extracción de la URL mediante reconciliación en Open Refine

Se ingresa la expresión cell.recon.match.id, la cual mediante la reconciliación extrae la URL

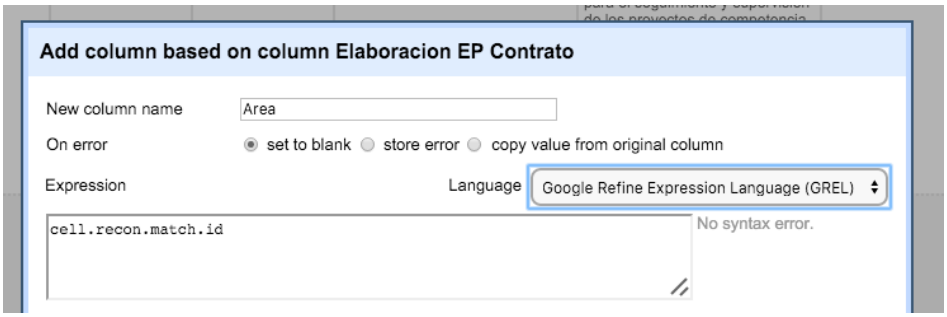


Figura 31. APLICACION de la expresión cell.recon.match.id en Open Refine

3.3.11 Consulta de Información

La consulta de información es una necesidad propia en las aplicaciones desarrolladas para la Web Semántica.

Una de las alternativas existentes para operar nuestra base de conocimiento es mediante un lenguaje de consulta. Actualmente existen varios lenguajes de consulta para ontologías RDF/XML, por ejemplo SPARQL y SQWRL. En nuestro caso de prueba para acceder a la base de conocimiento se utiliza el lenguaje de consulta SPARQL ya que ofrece una sintaxis de alto nivel, similar al lenguaje de consulta estructurado SQL, lo que hace que sea fácil su aprendizaje.

Para este trabajo, se plantea la utilización de una interfaz de consulta (GUI).

3.3.11.1 Interfaz de Formulación de Consultas (GUI)

Esta es una alternativa, la cual está destinada para usuarios finales que necesitan obtener información donde intervenga el razonamiento artificial mediante el uso de interfaz gráfica (GUI) y un lenguaje de consulta semántico (SPARQL).

Para nuestro caso puntual utilizaremos la herramienta virtuoso (OpenLink Software, 2012)

Figura 32. Interfaz Virtuoso Conductor

La cual nos facilita diferentes herramientas para la manipulación de la información contenida en nuestro esquema RDF. Para nuestro caso de prueba utilizaremos Virtuoso SPARQL Query Editor el cual nos permite hacer consultas utilizando SPARQL sobre nuestro esquema de datos.

Figura 33. Interfaz de Consulta Virtuoso SPARQL

Con el objetivo de validar si realmente los datos obtenidos con éxito realmente representan la información original (contenida en los documentos de texto semiestructurado), realizaremos algunas consultas de prueba y observaremos los resultados.

3.3.12 Publicación y Enlazado

Los archivos RDF producidos, han sido publicados en el servidor de pruebas donde está almacenado el prototipo, de manera que sirven como base para la gestión de datos enlazados, posibilitando que en el momento de pasar a producción sea posible su posterior consumo y visualización

Para realizar la publicación, se accede al prototipo mediante la url: http://localhost/prototipoTesis

s	p	o
http://www.co-ode.org/ontologies/ont.owl#ahorros	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#cedulaDeCiudadania	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#cedulaDeExtranjeria	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#contratista	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#corriente	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#funcionario	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#otro	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#otros	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://www.co-ode.org/ontologies/ont.owl#prestacionDeServicios	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/CDP	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/cargo	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/contrato	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/cuentaBancaria	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/persona	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/tipoContrato	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/tipoCuenta	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/tipoDocumento	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#Class
http://xmlns.com/foaf/0.1/supervisor	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/cargo	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/domicilio	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/elaboracionEP	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/fechacdp	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/idpersona	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/nocdp	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/nocontrato	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/nocuenta	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/nombre	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/objeto	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/plazoejecucion	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/tipocontrato	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/tipodecuenta	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/tipodocumento	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/valorTotal	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty
http://xmlns.com/foaf/0.1/valormensual	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/2002/07/owl#AnnotationProperty

Figura 34. Resultado de la consulta con las propiedades y objetos del esquema del caso de prueba

De manera que con este prototipo de aplicación se plantea en un escenario en el cual la base de conocimiento es usada para fines particulares aprovechando las características de la extracción de información y la publicación de datos a través de la web semántica.

3.4 Pruebas

Los resultados obtenidos al realizar la implementación permiten evaluar el método propuesto de manera que se pueda garantizar que los datos obtenidos realmente correspondan con la información contenida en los documentos de texto semiestructurados al interior de la entidad.

Para evaluar la propuesta se diseñaron 15 consultas en lenguaje SPARQL, las cuales deben permitir la correcta extracción de información, la cual debe corresponder con la información que fue cargada en la fase de poblado de la ontología.

La metodología para aplicar la evaluación de la propuesta consiste en ejecutar las consultas y luego verificar que los resultados obtenidos representen la información existente en el conjunto de documentos CONTRATOS, además de representar de manera correcta los datos requeridos por los parámetros de la consulta. De manera que si el resultado obtenido representa adecuadamente tanto la información real del documento, así como el concepto correcto en la ontología este resultado se interpretara como válido, en caso contrario se interpretara como invalido. **En el Anexo C se encuentran las 15 consultas que permitieron evaluar la propuesta.**

3.5 Resumen

En este capítulo se presentó el desarrollo e implementación del método propuesto, mediante la utilización de diversas herramientas, metodologías y estrategias, las cuales fueron claves para garantizar el éxito de la implementación del caso de prueba.

4. Conclusiones y Recomendaciones

4.1 Conclusiones

Actualmente la adopción de la Web Semántica se ha visto obstaculizada gracias a las brechas existentes entre las tecnologías utilizadas en la Web tradicional y las tecnologías necesarias para garantizar el apropiado funcionamiento de la Web Semántica.

Esto implica que las aplicaciones orientadas a la Web semántica sean prácticamente inexistentes, esto en gran medida se debe a la poca disponibilidad de datos enlazados; además los ya existentes carecen de las funcionalidades necesarias para garantizar la producción eficiente de datos enlazados.

En cuanto a este trabajo final de maestría, la principal motivación para su desarrollo fue la creciente necesidad de indexación de documentos la interior del SENA, lo que lleva a la necesidad de encontrar mecanismos que permitan representar la semántica de la información contenida en documentos de texto semiestructurado en un medio el cual pueda ser computacionalmente procesable.

Como resultado de este trabajo final de maestría se presenta un método para la extracción de información a partir de documentos semiestructurados producidos al interior del Servicio Nacional de Aprendizaje SENA, permitiendo su publicación, reutilización e intercambio a través de la web semántica, método el cual comprende una serie de fases y pasos que son desarrollados detalladamente, con el objetivo de obtener la información requerida; para esto fue necesario, la revisión y valoración de diversas herramientas que facilitasen el desarrollo de las actividades propuestas.

Además se pudo verificar que es posible representar la información contenida en documentos a un medio de representación semántico y éste a su vez puede emplearse para la publicación de información en un estándar como la Web Semántica.

4.2 Dificultades

La producción de documentos generados por instituciones de distinta índole crece constantemente, lo que dificulta el procesamiento de éstos altos volúmenes de información; además la información contenida en estos documentos posee un formato que sólo es comprensible a nivel humano, pero existen algunas estrategias para que la información de tales documentos sea mínimamente procesable/comprendida por computadoras.

Para nuestro caso particular se pudo verificar que lamentablemente la producción de este tipo documental al interior del SENA no sigue plenamente un estándar, sino que es común encontrar ligeras variaciones en la estructura del contenido, lo cual dificulta la extracción de información.

La extracción de información mediante la definición y utilización de patrones de texto puede resultar compleja debido a las variantes imprescindibles, como pueden ser signos de puntuación. Esto debido a que, es una tarea compleja para una maquina o herramienta de software hallar el significado de un elemento de texto que no concuerda con el patrón.

El formato de los documentos producidos al interior de la Entidad, no es homogéneo de manera que se presentan pequeñas variaciones en el contenido de los documentos, lo cual dificulta extraer Informacion de altos volúmenes de documentos siendo necesaria la creación validaciones adicionales.

Debido a que las letras acentuadas se encuentran excluidas en el conjunto de caracteres alfanuméricos, se hizo necesario implementar validaciones adicionales impactando los tiempos de desarrollo.

En la literatura actual realmente no existe un conjunto de buenas prácticas o recomendaciones para el diseño de expresiones regulares.

Los mecanismos existentes para la representación semántica de la información (RDF) limitan el escenario de prueba.

4.3 Limitaciones

- Es realmente complejo garantizar que se cubran todas las posibles variaciones al momento de realizar la búsqueda de patrones, para capturar la información de forma precisa.
- Actualmente es difícil encontrar Ontologias que cubran las necesidades particulares de un escenario en particular, por lo tanto una buena práctica es la reutilización de las ya existentes.
- Es de suma importancia que el formato de los documentos de origen tenga ciertas características, que faciliten las tareas necesarias para la para la extracción de texto plano.
- Es sumamente recomendable que el texto extraído, contenga un orden coherente.
- La presencia de ciertas estructuras (como tablas, graficas e imágenes producen resultados no esperados) lo que puede alterar la extracción correcta de información.
- El método propuesto sólo fue validado en un sólo escenario al interior del Servicio Nacional de Aprendizaje SENA.
- En caso de definir un nuevo escenario de prueba al interior de la Entidad será necesario crear nuevas reglas de inferencia.

4.4 Recomendaciones y Trabajo Futuro

En el caso hipotético de aumentar el alcance de este trabajo, será necesario que el uso actual de la ontología utilizado sea ampliamente extendido, y para esto será necesario garantizar la producción constante de datos enlazados,

Es necesario tener presente que al interior de la Entidad, existen gran cantidad de tipos

documentales que a futuro también requieren ser tratados, debido a que también contienen información vital para la toma de decisiones en el SENA, y aunque han quedado fuera del alcance, estos podrían ser incluidos a futuro; .De manera que al tener disponible una mayor cantidad de datos e información, sería posible realizar consultas con mayor complejidad impactando enormemente los procesos de gerencia estratégica al interior de la Entidad, abriendo la posibilidad de que mediante desarrollos adicionales sobre el prototipo puedan ser realizadas consultas que se ajusten a las necesidades de los usuarios finales.

Es deseable realizar un análisis sobre diversas entidades y organización públicas que permitan a futuro proponer una ontología de uso común.

Teniendo como referencia este trabajo es indispensable proponer buenas prácticas que permitan estandarizar la definición de expresiones regulares y así garantizar la correcta extracción de informació y un mejor desempeño.

A. Anexo: Implementaciones y Desarrollo

Se adjunta en CD, código fuente de los prototipos desarrollados.

B. Anexo: Conjunto de Datos Obtenidos

Se adjunta documentos PDF/A utilizados como insumo.

Se adjunta export de estructura BD en MySQL

C. Anexo: Conjunto de Consultas de Prueba

1. Consultar los contratistas existentes:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select *
from  <http://localhost/prototipoTesis/>
where
{
?persona foaf:idpersona ?idpersona
?persona foaf:nombre ?nombre
}
```

2. Consultar los contratos existentes:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select *
from  <http://localhost/prototipoTesis/>
where
{
?contrato foaf:nocontrato ?contrato
}
```

3. Consultar el objeto del contrato de la persona con documento = 1012432142

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?objeto
from  <http://localhost/prototipoTesis/>
where
{
?persona foaf:idpersona ?persona 1012432142
}
```

4. Consultar de manera ordenada las cuentas bancarias de los contratistas:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?nocuenta
from  <http://localhost/prototipoTesis/>
where
{
?persona foaf:idpersona ?idpersona
?persona foaf:nombre ?nombre
}
```

5. Consultar el valor total del contrato 12:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

Select ?valorTotal
from  <http://localhost/prototipoTesis/>
where
{
?contrato foaf:nocontrato 12
?contrato foaf:valorTotal ?contrato
}
```

6. Consultar todas las propiedades de nuestro URI:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?p ?o
{
       <http://localhost/prototipoTesis/>
       ?p ?o
}
```

7. Nombre de los contratistas cuyo valor mensual del contrato es menor a $3’000,000:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select *?nombre
from  <http://localhost/prototipoTesis/>
where
{
FILTER (?valormensual > 3000000)
}
```

8. Nombre de los contratistas cuyo valor mensual del contrato es mayor a $3’000,000 y menor a $6’000,0000

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select *?nombre
from  <http://localhost/prototipoTesis/>
where
{
FILTER (?valormensual > 3000000 && (?valormensual < 6000000)
}
```

9. Consultar contratos donde el supervisor es JOSE FERNANDO GALINDO:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?nocontrato
from  <http://localhost/prototipoTesis/>
where
{
?contrato foaf:nocontrato ?contrato
?contrato foaf:supervisor ?contrato “JOSE FERNANDO GALINDO”
}
```

10. Consultar Numero del CDP de los contratos:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?nocontrato, ?nocdp
from  <http://localhost/prototipoTesis/>
where
{
?nocdp foaf:CDP ?nocdp
}
```

11. Consultar diferentes tipos de cuenta bancaria:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select *
from  <http://localhost/prototipoTesis/>
where
{
?tipodecuenta foaf:tipoCuenta ?tipodecuenta
}
```

12. Consultar diferentes tipos de documento:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?tipodocumento
From  <http://localhost/prototipoTesis/>
Where
{
?tipoDocumento foaf:tipodocumento ?tipoDocumento
}
```

13. Consultar número de cuenta bancario del contratista LINA MARCELA AMAYA BAYONA:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?nocuenta Where { }
   values (?nombre)
{
   (“LINA MARCELA AMAYA BAYONA:”)
}
```

14. Consultar cuantos CDP distintos existen:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

select ?nocdp count(distinct ?nocdp)
where {
    ?nocdp foaf:CDP ?nocdp
    }
```

15. Consultar los tipo de contrato existentes:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

select ?tipocontrato
from  <http://localhost/prototipoTesis/>
where
{
?tipoContrato foaf:tipocontrato ?tipoContrato
}
```

5. Bibliografía

Adrian, B., Hees, J., Van Elst, L., & Dengel, A. (2009). iDocument: Using Ontologies for Extracting and Annotating Information from Unstructured Text. En Proceedings of the 32Nd Annual German Conference on Advances in Artificial Intelligence (pp. 249–256). Berlin, Heidelberg: Springer-Verlag. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1814110.1814147

Adrian, G. (2009). Handbook of Research on Emerging Rule-Based Languages and Technologies: Open Solutions and Approaches: Open Solutions and Approaches. IGI Global.

Altova. (2007). Enterprise Data Modeling Using XML Schema.

Angles, R., & Gutierrez, C. (2008). The Expressive Power of SPARQL. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 114–129). Springer. http://doi.org/10.1007/978-3-540-88564-1_8

Arroyo, & Otros. (2008). “La Educación y la Web Semántica”, 7.

Ben Mustapha, N., Zghal, H. B., Aufaure, M.-A., & ben Ghezala, H. (2010). Semantic Search Using Modular Ontology Learning and Case-based Reasoning. En Proceedings of the 2010 EDBT/ICDT Workshops (pp. 3:1–3:12). New York, NY, USA: ACM. http://doi.org/10.1145/1754239.1754243

Berners-Lee, T. (2000). Weaving the Web: the past, present and future of the World Wide Web by its inventor. London [u.a.]: Texere.

Bizer, C. (2011). Evolving the Web into a Global Data Space. En Proceedings of the 28th British National Conference on Advances in Databases (pp. 1–1). Berlin, Heidelberg: Springer-Verlag. Recuperado a partir de http://dl.acm.org/citation.cfm?id=2075914.2075915

Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., & Völker, J. (2013). Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis. En H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, … K. Janowicz (Eds.), The Semantic Web – ISWC 2013 (pp. 17–32). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-41338-4_2

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., & Hellmann, S. (2009). DBpedia - A Crystallization Point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web, 7(3), 154–165. http://doi.org/10.1016/j.websem.2009.07.002

Blackwell. (2004). The Blackwell Guide to the Philosophy of Computing and Information.

Borges, K. A. V., Jr, C. A. D., Laender, A. H. F., & Medeiros, C. B. (2010). Ontology Driven Discovery of Geospatial Evidence in Web Pages. GeoInformatica, 15(4), 609–631. http://doi.org/10.1007/s10707-010-0118-z

BOSAK, J., & Tim Bray. (1999). XML and The Second Generation Web.

Broekstra, J., Kampman, A., & Harmelen, F. van. (2002). Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. En Proceedings of the First International Semantic Web Conference on The Semantic Web (pp. 54–68). London, UK, UK: Springer-Verlag. Recuperado a partir de http://dl.acm.org/citation.cfm?id=646996.711426

Buil-Aranda, C., Hogan, A., Umbrich, J., & Vandenbussche, P.-Y. (2013). SPARQL Web-Querying Infrastructure: Ready for Action? En H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, … K. Janowicz (Eds.), The Semantic Web – ISWC 2013 (pp.

277–293). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-41338-4_18

Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., & Racioppa, S. (2008). Ontology-based information extraction and integration from heterogeneous data sources. International Journal of Human-Computer Studies, 66(11), 759–788. http://doi.org/10.1016/j.ijhcs.2008.07.007

Burstein, M. (2004, noviembre). OWL-S: Semantic Markup for Web Services. Recuperado a partir de http://www.w3.org/Submission/OWL-S/.

Calvo L., J. M. (2006). Web Semántica ó Web 3.0 La Web con significado.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., & Wilkinson, K. (2004). Jena: Implementing the Semantic Web Recommendations. En Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters (pp. 74–83). New York, NY, USA: ACM. http://doi.org/10.1145/1013367.1013381

Chen, R.-C., Bau, C.-T., & Huang, Y.-H. (2010). Development of anti-diabetic drugs ontology for guideline-based clinical drugs recommend system using OWL and SWRL (pp. 1–6). IEEE. http://doi.org/10.1109/FUZZY.2010.5584139

Chen Shao-fei, H. Y., Tian-zhu, X. L., & Yang Wen-zhu. (2003). Evolution of Information Extraction Techniques on the Web. In Journal of Hebei University, volume 23, pages 106–111, 2003.

Damljanovic, D., Amardeilh, F., & Bontcheva, K. (2009). CA Manager Framework: Creating Customised Workflows for Ontology Population and Semantic Annotation. En Proceedings of the Fifth International Conference on Knowledge Capture (pp. 177–178). New York, NY, USA: ACM. http://doi.org/10.1145/1597735.1597770

Daniel Krech. (2012). RDFLib. Recuperado a partir de https://github.com/gjhiggins/rdflib

Desire Consortium. (2000, agosto 7). Desire: Development of a European Service for Information on Research and Education.

Dimitrova, V., Denaux, R., Hart, G., Dolbear, C., Holt, I., & Cohn, A. G. (2008). Involving Domain Experts in Authoring OWL Ontologies. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC

2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 1–16). Springer. http://doi.org/10.1007/978-3-540-88564-1_1

Dublin Core. (2012, julio). Dublin Core. The Dublin Core Metadata Initiative. Recuperado a partir de http://dublincore.org

Friend of a Friend. (2012). Recuperado a partir de Friend of a Friend

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. En Proceedings of the 16th Conference on Computational Linguistics - Volume 1 (pp. 466–471). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/992628.992709

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), 199–220. http://doi.org/10.1006/knac.1993.1008

Han, L., Finin, T., Parr, C. S., Sachs, J., & Joshi, A. (2008). RDF123: From Spreadsheets to RDF. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 451–466). Springer. http://doi.org/10.1007/978-3-540-88564-1_29

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology. Recuperado a partir de http://linkeddatabook.com/editions/ 1.0

Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., & Tyson, M. (1997). FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. Finite-state language processing, 383.

Horrocks, I., Patel-Schneider, P. F., & Van Harmelen, F. (2003). From SHIQ and RDF to OWL: The Making of a Web Ontology Language. Web semantics: science, services and agents on the World Wide Web, 1(1), 7–26.

Huang, Z., & Harmelen, F. van. (2008). Using Semantic Distances for Reasoning with Inconsistent Ontologies. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 178–194). Springer. http://doi.org/10.1007/978-3-540-88564-1_12

Hu, X., Lin, T. Y., Song, I.-Y., Lin, X., Yoo, I., Lechner, M., & Song, M. (2004). Ontology-Based Scalable and Portable Information Extraction System to Extract Biological Knowledge from Huge Collection of Biomedical Web Documents. En Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (pp. 77–83). Washington, DC, USA: IEEE Computer Society. http://doi.org/10.1109/WI.2004.109

Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., & Hitzler, P. (2013). A Linked Data Driven and Semantically Enabled Journal Portal for Scientometrics. En H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, … K. Janowicz (Eds.), The Semantic Web – ISWC 2013 (pp. 114–129). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-41338-4_8

Hyland. (1991). OnBase. Recuperado a partir de https://www.onbase.com/

Ian Horrocks. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML.

(ISO). (1996). Extended BNF (Syntactic Metalanguage). Recuperado a partir de http://www.iso.org/iso/catalogue_ detail.htm?csnumber=26153.

Johan Hjelm. (2001). Creating the Semantic Web with RDF: Professional Developer's Guide. John Wiley & Sons, Inc.

Junghoo Cho, S. R. (2012). A Fast Regular Expression Indexing Engine.

Jurafsky, D., & Martin, J. H. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

Karvounarakis, G., | S., Christophides, V., Plexousakis, D., & Scholl, M. (2002). RQL: A Declarative Query Language for RDF. En Proceedings of the 11th international

conference on World Wide Web (pp. 592–603). ACM. Recuperado a partir de http://dl.acm.org/citation.cfm?id=511524

Khadir, M. T., Djeddai, A., & Djeddi, W. (2011). XMap: A Novel Semantic Approach for Alignment of OWL-Full Ontologies Based on Semantic Relationship Using WordNet. En 2011 Fourth International Symposium on Innovation in Information Communication Technology (ISIICT) (pp. 13–18). http://doi.org/10.1109/ISIICT.2011.6149595

Kriglstein, S., & Wallner, G. (2010). Knoocks - A Visualization Approach for OWL Lite Ontologies. En 2010 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS) (pp. 950–955). http://doi.org/10.1109/CISIS.2010.55

Lavelli, A., Califf, M. E., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., … Ireson, N. (2008). Evaluation of Machine Learning Based Information Extraction Algorithms: Criticisms and Recommendations. Language Resources and Evaluation, 42(4), 361–393. http://doi.org/10.1007/s10579-008-9079-3

Liu, G., Wang, Y., & Wu, C. (2010). Research and Application of Geological Hazard Domain Ontology. En 2010 18th International Conference on Geoinformatics (pp. 1–6). http://doi.org/10.1109/GEOINFORMATICS.2010.5567498

Liu, S., Yang, Y., Xie, G. T., Wang, C., Cao, F., Santos, C. D., … Colgrave, J. (2008). Supporting Ontology-Based Dynamic Property and Classification in WebSphere Metadata Server. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 861–874). Springer. http://doi.org/10.1007/978-3-540-88564-1_56

LOD2. (2013). Creating Knowledge Out of Interlinked Data. Recuperado a partir de http://stack.lod2.eu/blog/

Luna, G., A, J., Torres Pardo, D., & Ovalle, D. A. (2007). SABIOS: Una Aplicación de la Web Semántica Para la Gestión de Documentos Digitales. Revista Interamericana de Bibliotecología, 30(1), 51–71.

Martínez Salinas, O. (2012, Diciembre). Modelado Semántico de Documentos con Estructura Definida.

Martin Kaltenböck, & Florian Bauer. (2012). Linked Open Data: The Essentials A Quick Start Guide for Decision Makers.

McIlraith, S. A., Plexousakis, D., & Harmelen, F. van. (2004). The Semantic Web - ISWC 2004: Third International Semantic Web Conference. Springer.

Miller, L., Seaborne, A., & Reggiori, A. (2002). Three Implementations of SquishQL, a Simple RDF Query Language. En I. Horrocks & J. Hendler (Eds.), The Semantic Web — ISWC 2002 (pp. 423–435). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/3-540-48005-6_36

Novotny, R., Vojtas, P., & Maruscak, D. (2009). Information Extraction from Web Pages. En IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT '09 (Vol. 3, pp. 121–124). http://doi.org/10.1109/WI-IAT.2009.245

Noy, N. F., Griffith, N., & Musen, M. A. (2008). Collecting Community-Based Mappings in an Ontology Repository. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 371–386). Springer. http://doi.org/10.1007/978-3-540-88564-1_24

Noy, N. F., & McGuinness, D. L. (2012). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, octubre 2012. Recuperado a partir de http://protege.stanford. edu/publications/ontology_development/ontology101-noy-mcguinness.html

OASIS. (2009). UIMA (Unstructured Information Management applications).

OpenLink Software. (2012). OpenLink Virtuoso. Recuperado a partir de https://www.w3.org/2001/sw/wiki/

OpenRefine. (2012).

Ora Lassila. (1997). Introduction to RDF Metadata. Recuperado a partir de http://www.w3.org/TR/ NOTE-rdf-simple-intro.

PDF/A Competence Center. (2011). PDF/A. Recuperado a partir de http://www.pdfa.org/wp-content/uploads/2011/07/pdfa-flyer-esp.pdf

Peis, E., Herrera-Viedma, E., Hassan, Y., & Herrera, J. C. (2003). Análisis de la Web Semántica: Estado Actual y Requisitos Futuros [Journal article (Print/Paginated)]. Recuperado el 19 de noviembre de 2015, a partir de http://eprints.rclis.org/11446/

Pérez, J., Arenas, M., & Gutierrez, C. (2008). nSPARQL: A Navigational Language for RDF. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 66–81). Springer. http://doi.org/10.1007/978-3-540-88564-1_5

Perry, M., Jain, P., & Sheth, A. P. (2011). SPARQL-ST: Extending SPARQL to Support Spatiotemporal Queries. En N. Ashish & A. P. Sheth (Eds.), Geospatial Semantics and the Semantic Web (pp. 61–86). Springer US. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-1-4419-9446-2_3

Piedra, N., Tovar, E., Colomo-Palacios, R., Lopez-Vargas, J., & Chicaiza, J.A. (2014). Consuming and Producing Linked Open Data: The case of OpenCourseWare.

Polleres, A., Scharffe, F., & Schindlauer, R. (2007). SPARQL++ for Mapping Between RDF Vocabularies. En Proceedings of the 2007 OTM Confederated International Conference on On the Move to Meaningful Internet Systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part I (pp. 878–896). Berlin, Heidelberg: Springer-Verlag. Recuperado a partir de http://dl.acm.org/citation.cfm?id=1784607.1784685

Quiroga, G. B. B. (2015, Agosto). Prototipo de un sistema de integración de recursos científicos, diseñado para su funcionamiento en el espacio de los datos abiertos enlazados para mejorar la colaboración, la eficiencia y promover la innovación en Colombia. Recuperado a partir de http://www.bdigital.unal.edu.co/50580/

RAP - RDF API for PHP. (2002). Recuperado a partir de http://wifo5-03.informatik.uni-mannheim.de/bizer/rdfapi/

Richard y Heath-Tom Bizer, C. y C. (2007). How to Publish Linked Data on the Web. Recuperado a partir de http://www4.wiwiss.fu-berlin.de/bizer/pub/ LinkedDataTutorial

Rodríguez, C. M., Montaño, W. C., & Martínez, J. M. (2010). Razonadores semánticos: un estado del arte. Ingenium, (21). Recuperado a partir de http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=01247492&AN=54490576&h=2obMoGLM19JCg3SlWE2SkH%2BbkqbTnE7tfgI7RLbKLUwJM4lNlozlrJxcnnFOUqjCFEqtqtve1e1mSecR9UPRxg%3D%3D&crl=c

Rodríguez Mendez, E. M. (1999). RDF: un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio. En Les biblioteques i els centres de documentació al segle XXI: peça clau de la societat de la informació (pp. 487–498). Recuperado a partir de http://www.researchgate.net/profile/Eva_Mendez/publication/28809837_RDF__Un_modelo_de_metadatos_flexible_para_las_bibliotecas_digitales_del_prximo_milenio/links/54629b360cf2c0c6aec1b5b9.pdf

Ropero Rodríguez, J. (2009). Método general de Extracción de Información basado en el uso de Lógica Borrosa. Aplicación en portales web. Recuperado a partir de http://www.dte.us.es/personal/jropero/TesisV2p0.pdf

Sáez Guerrero, M. (2009). Diseño de un Sistema de Extracción de Información de Artículos de Wikipedia. Recuperado a partir de http://e-archivo.uc3m.es/handle/10016/5874

Schmidt, M., Hornung, T., Küchlin, N., Lausen, G., & Pinkel, C. (2008). An Experimental Comparison of RDF Data Management Approaches in a SPARQL Benchmark Scenario. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 82–97). Springer. http://doi.org/10.1007/978-3-540-88564-1_6

sebastien malot. (2013). pdfparser. Recuperado a partir de http://www.pdfparser.org/

Semantic Web for Research Communities (SWRC). (2012). Recuperado a partir de http://ontoware. org/swrc

Sesame. (2012). Recuperado a partir de http://www.openrdf.org

Sheth, A. P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T. W., & Thirunarayan, K. (Eds.). (2008). The Semantic Web - ISWC 2008, 7th International Semantic Web Conference (Vol. 5318). Springer.

Sirin, E., & Parsia, B. (2007). SPARQL-DL: SPARQL Query for OWL-DL. En OWLED (Vol. 258). Recuperado a partir de http://owl1-1.googlecode.com/svn-history/r603/trunk/www.webont.org/owled/2007/PapersPDF/submission_23.pdf

Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., & Oberle, D. (2005). The SWRC ontology–semantic web for research communities. En Progress in Artificial Intelligence (pp. 218–231). Springer. Recuperado a partir de http://link.springer.com/chapter/10.1007/11595014_22

The Protege Project. (2000). Recuperado a partir de http://protege.stanford.edu

Thompson, K. (1968). Programming Techniques: Regular expression search algorithm. Communications of the ACM, 11(6), 419–422. http://doi.org/10.1145/363347.363387

Vargas-Vera, M., Moreale, E., Stutt, A., Motta, E., & Ciravegna, F. (2007). MnM: Semi-Automatic Ontology Population from Text. En R. Sharman, R. Kishore, & R. Ramesh (Eds.), Ontologies (pp. 373–402). Springer US. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-0-387-37022-4_13

Volker Haars lev. (2004). Querying the Semantic Web with Racer + nRQL.

W3C. (2004). RDF Schema. Recuperado a partir de http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

W3C Working Group. (2006). Guía Breve de Linked Data. Recuperado a partir de http://www.w3c.es/Divulgacion/GuiasBreves/LinkedData

Wang, H., Zhai, S., & Fan, L. (2009). Query for Semantic Web Services Using SPARQL-DL. En Second International Symposium on Knowledge Acquisition and Modeling, 2009. KAM ’09 (Vol. 1, pp. 367–370). http://doi.org/10.1109/KAM.2009.198

Wang, S., Englebienne, G., & Schlobach, S. (2008). Learning Concept Mappings from Instance Similarity. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 339–355). Springer. http://doi.org/10.1007/978-3-540-88564-1_22

Wick, M. (2012). GeoNames: Geographical Database. Recuperado a partir de http://www.geonames.org

World Wide Web Consortium. (2012a, agosto). A Direct Mapping of Relational Data to RDF. Recuperado a partir de http://www.w3.org/TR/2012/REC-rdb-direct-mapping-20120927/.

World Wide Web Consortium. (2012b, septiembre). RDF/XML Syntax Specification. Recuperado a partir de http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

World Wide Web Consortium. (2012c, octubre). OWL 1.1 Web Ontology Language: Structural Specification and Functional-Style Syn- tax. Recuperado a partir de http://www.w3.org/TR/2008/ WD-owl11-syntax-20080108/.

World Wide Web Consortium. (2012d, octubre). OWL Web Ontology Language. Recuperado a partir de http: //www.w3.org/TR/2004/REC-owl-features-20040210/.

World Wide Web Consortium. (2012e, octubre). RDF Vocabulary Description Language 1.0: RDF Schema. Recuperado a partir de http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

Wu, G., Li, J., Feng, L., & Wang, K. (2008). Identifying Potentially Important Concepts and Relations in an Ontology. En A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, & K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26-30, 2008. Proceedings (Vol. 5318, pp. 33–49). Springer. http://doi.org/10.1007/978-3-540-88564-1_3

Wu, J., Ilyas, I., & Weddell, G. (2011). A Study of Ontology Based Query Expansion. Technical report CS-2011-04, University of Waterloo. Recuperado a partir de https://cs.uwaterloo.ca/research/tr/2011/CS-2011-04.pdf

Yu, C.-H., Groza, T., & Hunter, J. (2013). Reasoning on Crowd-Sourced Semantic Annotations to Facilitate Cataloguing of 3D Artefacts in the Cultural Heritage Domain. En H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. X. Parreira, … K. Janowicz (Eds.), The Semantic Web – ISWC 2013 (pp. 228–243). Springer Berlin Heidelberg. Recuperado a partir de http://link.springer.com/chapter/10.1007/978-3-642-41338-4_15

Zhou, J., Ma, L., Liu, Q., Zhang, L., Yu, Y., & Pan, Y. (2006). Minerva: A Scalable OWL Ontology Storage and Inference System. En R. Mizoguchi, Z. Shi, & F. Giunchiglia (Eds.), The Semantic Web – ASWC 2006 (Vol. 4185, pp. 429–443). Springer Berlin Heidelberg. Recuperado a partir de http://dx.doi.org/10.1007/11836025_42