**UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA**

*La Universidad Católica de Loja*

**ÁREA TÉCNICA**

TITULACIONES DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

**Desarrollo de Servicios Web para el proceso de Enlace y Enriquecimiento de Datos   Enlazados**

TRABAJO DE FIN DE TITULACIÓN

**Autor:** Montaño Sozoranga, Wilmer Fabricio

**Director:** Piedra Pullaguari, Nelson Oswaldo, Ing

Loja - Ecuador

2014

Contenido

[Indice de FIguras 5](#_Toc393117291)

[Índice de tablas 6](#_Toc393117292)

[CAPITULO I: MARCO TEÓRICO 7](#_Toc393117293)

[1.1. Datos Enlazados 8](#_Toc393117294)

[1.1.1. Introducción 8](#_Toc393117295)

[1.1.2. Principios de Datos Enlazados 8](#_Toc393117296)

[1.1.3. Tecnologías 9](#_Toc393117297)

[1.1.3.1. URI 9](#_Toc393117298)

[1.1.3.2. RDF 12](#_Toc393117299)

[1.1.3.3. SPARQL Query Language for RDF 20](#_Toc393117300)

[1.1.4. Acerca de DBpedia 21](#_Toc393117301)

[1.1.4.1. Framework extracción 21](#_Toc393117302)

[1.1.4.2. DBpedia Dataset 22](#_Toc393117303)

[1.1.4.3. Acceso a DBpedia Dataset 22](#_Toc393117304)

[1.2. Procesamiento de Lenguaje Natural (PLN) 23](#_Toc393117305)

[1.2.1. Introducción 23](#_Toc393117306)

[1.2.2. Part of Speech Tagger 24](#_Toc393117307)

[1.2.3. Chunking 24](#_Toc393117308)

[1.2.4. Desambiguación 25](#_Toc393117309)

[1.2.4.1. Métodos basados en el conociendo. 25](#_Toc393117310)

[1.2.4.2. Métodos supervisados 26](#_Toc393117311)

[1.2.4.3. Métodos semi-supervisados 26](#_Toc393117312)

[1.2.4.4. Métodos sin supervisión 27](#_Toc393117313)

[1.3. Servicios Web 27](#_Toc393117314)

[1.3.1. Introducción 27](#_Toc393117315)

[1.3.2. Tipos de Web Services 27](#_Toc393117316)

[1.3.2.1. SOAP AND THE WS-\* STACK 27](#_Toc393117317)

[1.3.2.2. REST 28](#_Toc393117318)

[1.3.2.3. Hipermedia 29](#_Toc393117319)

[1.3.2.4. Recursos y representaciones 29](#_Toc393117320)

[1.3.2.5. URI y relación con los recursos 30](#_Toc393117321)

[1.4. Trabajos relacionados 30](#_Toc393117322)

[CAPITULO 2: PROBLEMÁTICA 31](#_Toc393117323)

[2.1. Estado actual 32](#_Toc393117324)

[2.2. Justificación 32](#_Toc393117325)

[2.3. Objetivos 33](#_Toc393117326)

[CAPITULO 3: Solución 34](#_Toc393117327)

[3.1. Propuesta 35](#_Toc393117329)

[3.2. Arquitectura 36](#_Toc393117330)

[3.2.1. Lógica de negocios (Aplicación) 37](#_Toc393117331)

[3.2.2. Dataset Local con recursos de DBpedia y base de conocimientos 37](#_Toc393117332)

[3.2.3. Frontal 38](#_Toc393117333)

[3.3. Metodología 38](#_Toc393117334)

[3.4. Riesgos 38](#_Toc393117335)

[3.5. Módulos 38](#_Toc393117336)

[3.5.1. Validación de Texto 38](#_Toc393117337)

[3.5.2. Módulo de etiquetado 38](#_Toc393117338)

[3.5.3. Extracción de Entidades y Keywords 40](#_Toc393117339)

[3.5.4. Desambiguación y Enlace 41](#_Toc393117340)

[3.6. Implementación 42](#_Toc393117341)

[CAPITULO 4 43](#_Toc393117342)

[4.1. Validación 43](#_Toc393117344)

[4.1.1. Comparación con servicios similares 43](#_Toc393117345)

[4.2. Pruebas 43](#_Toc393117346)

[CAPITULO 5 43](#_Toc393117347)

[5.1. Conclusiones y Recomendaciones 43](#_Toc393117349)

[Bibliografía 44](#_Toc393117350)

Indice de FIguras

[Figura 1: Relación entre URI, URL y URN. Fuente: propio 11](#_Toc393116457)

[Figura 2: Estructura de la URI (Scheme URI). Fuente: (Albahari & Albahari, 2012) 11](#_Toc393116458)

[Figura 3. RDF 1.0 y 1.1 formatos de serialización. 13](#_Toc393116459)

[Figura 4: RDF con dos nodos (Subject y Object) y una conexión entre ellos (Predicado). 13](#_Toc393116460)

[Figura 5: Representación gráfica RDF 22](#_Toc393116461)

[Figura 6. Arquitectura de provisión de Datos de Dbpedai. 23](#_Toc393116462)

[Figura 7: Ejemplo POS Tag y Chunking 25](#_Toc393116463)

[Figura 8. Lógica propuesta para la Aplicación; 35](#_Toc393116464)

[Figura 9. Arquitectura. 36](#_Toc393116465)

[Figura 10. Diagrama de secuencias del módulo de Etiquetado 39](#_Toc393116466)

[Figura 11. Diagrama de secuencias del módulo de Extracción de Entidades y Keywords 40](#_Toc393116467)

[Figura 12. Diagrama de secuencias módulo de Desambiguación y Enlace 41](#_Toc393116468)

Índice de tablas

[Tabla 1. Ejemplos de Prefijos de Espacios de Nombres e IRIs 14](#_Toc393117110)

[Tabla 3. Ejemplo N-triple 15](#_Toc393117111)

[Tabla 4. Ejemplo Turtle 16](#_Toc393117112)

[Tabla 5. Ejemplo TriG 16](#_Toc393117113)

[Tabla 6. Ejemplo N-Quads 17](#_Toc393117114)

[Tabla 7. Ejemplo JSON-LD 18](#_Toc393117115)

[Tabla 8. Ejemplo RDFa 18](#_Toc393117116)

[Tabla 9. Ejemplo RDF/XML 19](#_Toc393117117)

[Tabla 10. Resultado consulta SPARQL 21](#_Toc393117118)

CAPITULO I: MARCO TEÓRICO

* 1. Datos Enlazados
     1. Introducción

En sus inicios la web en su primera versión 1.0, donde web era rígida en cuanto a la entrega de información, además de poco actualizada, convertía al visitante de un sitio web un simple lector, restringido de cualquier interacción. Se puede decir que la web no era más que paginas enlazadas mediante hipervínculos.

La web que siempre está creciendo y evolucionado, alcanza su versión conocida como la web 2.0 en donde usuario juega el papel más importante, es quien evalúa, puede calificar, compartir, rectificar, alimentar y subir su propia información a la web. Esto producto de le aparición de nuevas tecnologías y estandarización[[1]](#footnote-1).

Los datos relacionados llegan para dar forma a la siguiente versiona de la web, la web semántica de W3C[[2]](#footnote-2) nos dice : “Linked Data se refiere a la utilización delas mejores prácticas para publicación, estructuración de los datos en la web, de tal forma que puedan ser enlazados entre sí, utilizando tecnología propias de web semántica como RDF, OCW, SPARQL, etc.”

Se refiere en si a la estructura de la de la siguiente generación de la web, como es la web semántica, que en si busca que la información que se en publica en internet pueda no solo ser entendida por seres humanos sino también por la maquinas que navegan en la web. En donde a partir de un dato podemos descubrir otros datos por sus relaciones.

* + 1. Principios de Datos Enlazados

Tim Berners Lee en su publicación Linked Data - Design Issues (Berners-Lee, Linked Data - Design Issues, 2006)describe cuatro reglas base para la publicación de datos enlazados:

1. Usar URIs como nombre de las cosas
2. Usar URIs HTTP par que esas cosas puedan ser referenciadas
3. Representar los datos en RDF y SPARQL como lenguaje de consulta
4. Incluir enlaces hacia otra cosas, para descubrir más cosas

La utilizar de estas reglas para la publicación de datos, nos permite que estos por las características propias de las tecnologías sobre las cuales se construyen como:

* Las cosas que nombremos por URIs son inequívocas y estos serán recursos.
* Los detalles o atributos y las relaciones de los datos van a estar descritos y estructurados en formato RDF
* Podremos acceder o realizar consultas sobre estos mediante SPARQL
* La cosas que publiquemos estarán relacionados
  + 1. Tecnologías
       1. URI

URI (Uniform Resource Identifier) ha sido desarrollado por el [IETF](http://www.ietf.org/)[[3]](#footnote-3) y pretende crear un sistema mundial para identificar recursos de todo tipo en la web: documentos, imágenes, programas, servicios, correos electrónicos, etc. Este método combina URNs y URLs, esto es, nombres/direcciones. Se trata de identificar los documentos mediante una secuencia de sintaxis controlada que identifica cada documento de una forma única. (...). Los URIs hacen posible encontrar los recursos bajo una gran variedad de esquemas definidos y métodos de acceso tales como HTTP, FTP, Gopher, news, telnet o correos electrónicos localizables siempre de la misma manera, ya que a un mismo documento se puede acceder desde distintos protocolos. Ya se han establecido una serie de schemes o esquemas direccionados. Los esquemas definidos URI coinciden con los protocolos más usados de Internet. (Lapuente, 2013)

En (Berners-Lee, 2005) el RFC[[4]](#footnote-4) que trata sobre este tema nos dice lo siguiente: “Un identificador uniforme de recursos (URI) proporciona un medio simple y extensible para la identificación de un recurso”. Es base a los términos que lo conforman se explica como:

“Uniforme: Uniformidad ofrece varios beneficios. Permite diferentes tipos de identificadores de recursos que se utilizarán en el mismo contexto, aun cuando los mecanismos utilizados para acceder a esos recursos pueden ser diferentes. Permite la interpretación semántica uniforme de convenciones sintácticas comunes a través de diferentes tipos de identificadores de recursos (…).

Recurso: Esta especificación no limita el alcance de lo que podría ser un recurso; más bien, el término "recurso" se utiliza en un sentido general de lo que pudiera ser identificado por un URI. Ejemplos conocidos incluyen un documento electrónico, una imagen, una fuente de información con un propósito consistente (por ejemplo, "parte meteorológico de hoy para Los Ángeles"), un servicio (por ejemplo, una puerta de enlace HTTP a SMS), y una colección de otros recursos. Un recurso no es necesariamente accesible a través de Internet; por ejemplo, los seres humanos, las empresas y los libros encuadernados en una biblioteca también pueden ser recurso (…).

Identificador: Un identificador encarna la información necesaria para distinguir lo que se identificó a partir de todas las otras cosas dentro de su ámbito de aplicación de la identificación. Nuestro uso de los términos "identificar" y "identificación" se refieren a este fin de distinguir un recurso de todos los demás recursos, independientemente de cómo se logra ese propósito (ejemplos, nombre, dirección, o el contexto). Estos términos no deben confundirse con la presunción de que un identificador define o encarna la identidad de lo que se hace referencia, aunque esto puede ser el caso de algunos identificadores. Tampoco debe asumirse que un sistema que utiliza los URI tendrá acceso al recurso identificado: en muchos casos, los URI se utilizan para referirse a los recursos sin ninguna intención de que se puede acceder. Del mismo modo, "un" recurso identificado podría no ser singular en la naturaleza (ejemplo, un recurso puede ser un conjunto con nombre o una asignación que varía con el tiempo).” (Berners-Lee, 2005)

Suele existir confusión entre URI, URN y URL, que las describimos a continuación de acuerdo a (Berners-Lee, 2005): “El término "Uniform Resource Locator" (URL[[5]](#footnote-5)) se refiere al subconjunto de URIs que, además de la identificación de un recurso, proporcionan los medios para localizar el recurso mediante la descripción de su mecanismo de acceso primario (por ejemplo, su red de "ubicación"). El término "Uniform Resource Name" (URN[[6]](#footnote-6)) se ha utilizado históricamente para referirse tanto a los URI en el marco del esquema de "URN" [RFC2141], que son necesarios para permanecer globalmente único y persistente, incluso cuando el recurso deja de existir o no está disponible, y a cualquier otro URI con las propiedades de un nombre”. La diferencia en grafico de estos conceptos se la puede ver en el Figura 1.

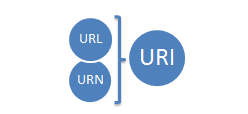


Figura 1: Relación entre URI, URL y URN. Fuente: propio

La estructura de una URI es explicada en el Figura 1 de (Albahari & Albahari, 2012), donde vemos los componentes: Scheme, Authority (Host, Port), PathAndQuery (AbsolutyPath, Query y Fragment), UserInfo.

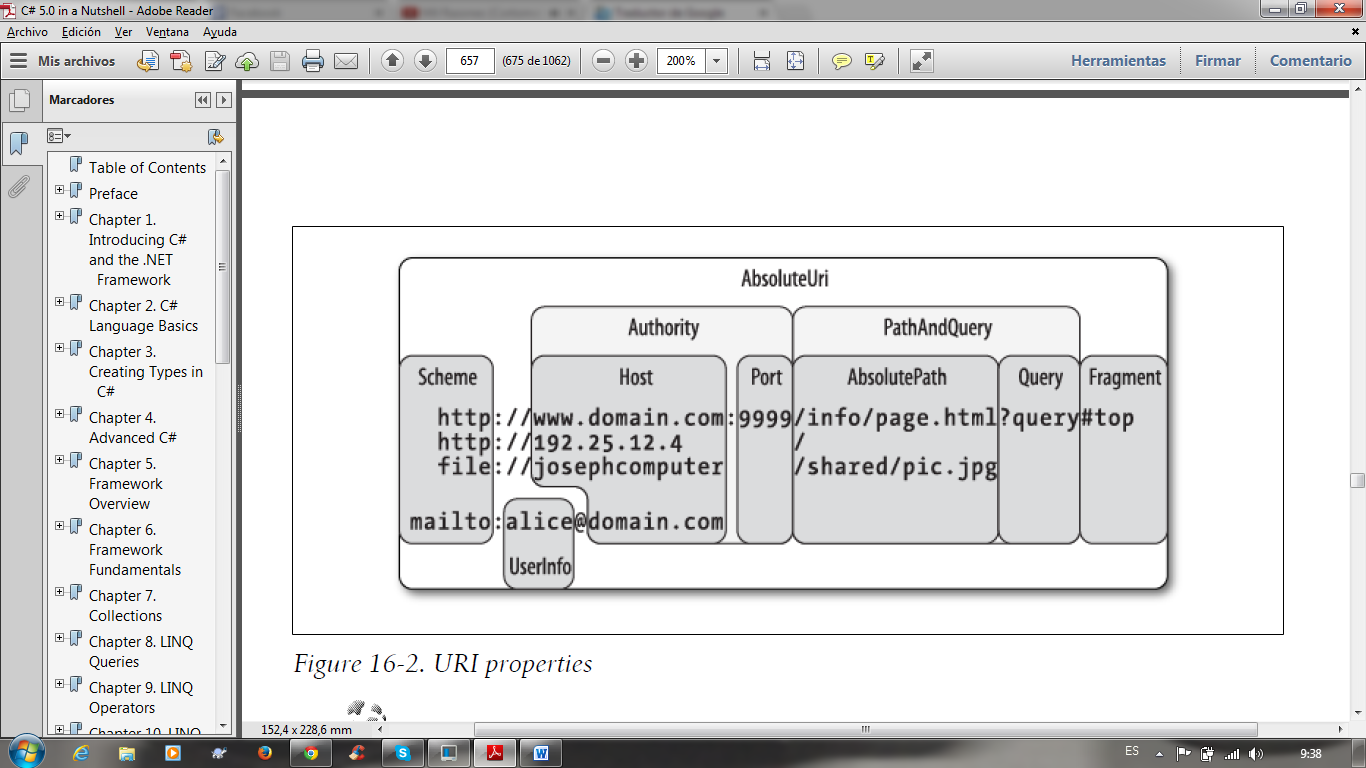


Figura 2: Estructura de la URI (Scheme URI). Fuente: (Albahari & Albahari, 2012)

(Lapuente, 2013) recoge alguna definiciones sobre las URIs de Architecture of the World Wide Web, Volume One especificación publicada por W3C, que exponemos acountinuacion:

* URI: Acrónimo para Uniform Resource Identificer.
* URI aliases: dos o más URIs que son -carácter por carácter-, diferentes, pero que identifican el mismo recurso.
* URI overloading: uso del mismo URI para referirse a más de un recurso en el contexto de los protocolos y formatos de la Web.
* URI ownership: la relación entre el agente que asigna y el URI que es definido por un esquema URI.
* URI persistence: la expectación social que desde hace algún tiempo identifica URI a un recurso particular, este podría continuar indefinidamente para referirse al recurso.
* URI reference: un apunte operacional para un URI.
* Uniform Resource Identifier (URI): un identificador global en el contexto de la World Wide Web.
* Namespace document: el recurso de información identificado por un [namespace](http://www.hipertexto.info/documentos/namespaces.htm) URI en [XML](http://www.hipertexto.info/documentos/xml.htm).
* Link o enlace: una relación entre dos recursos cuando un recurso (representación) se refiera al otro recurso mediante el significado de un URI.
  + - 1. RDF

El Marco de Descripción de Recursos (RDF) es un lenguaje para representar la información acerca de los recursos en la World Wide Web. En particular, se pretende para la representación de metadatos sobre recursos web, como el título, el autor y la fecha de modificación de una página Web, los derechos de autor y la información de licencia de un documento Web, o el calendario de disponibilidad de algún recurso compartido. Sin embargo, al generalizar el concepto de un "recurso de la Web", RDF puede también ser utilizado para representar la información acerca de las cosas que se pueden identificar en la red, incluso cuando no pueden recuperarse directamente en la Web. Los ejemplos incluyen información acerca de productos disponibles de servicios en línea de la compra (por ejemplo, información acerca de las especificaciones, los precios y la disponibilidad), o la descripción de las preferencias del usuario Web para la entrega de información. (McBride, 2004)

Desde la primera publicación de RDF en su versión 1.0 en febrero del 2004 han pasado 10 años para que en la primera mitad del 2014 poder conocer la versión 1.1, las novedades de esta nueva versión y las diferencias entre las versiones podemos encontrarlas en *What’s New in RDF 1.1*[[7]](#footnote-7)*,* entre las cuales se resalta dos, la primera que ahora se utiliza IRIs en lugar de URIs para identificar a los recursos y segundo nuevos formatos se serialización, este último lo visualizamos en la Figura 3.

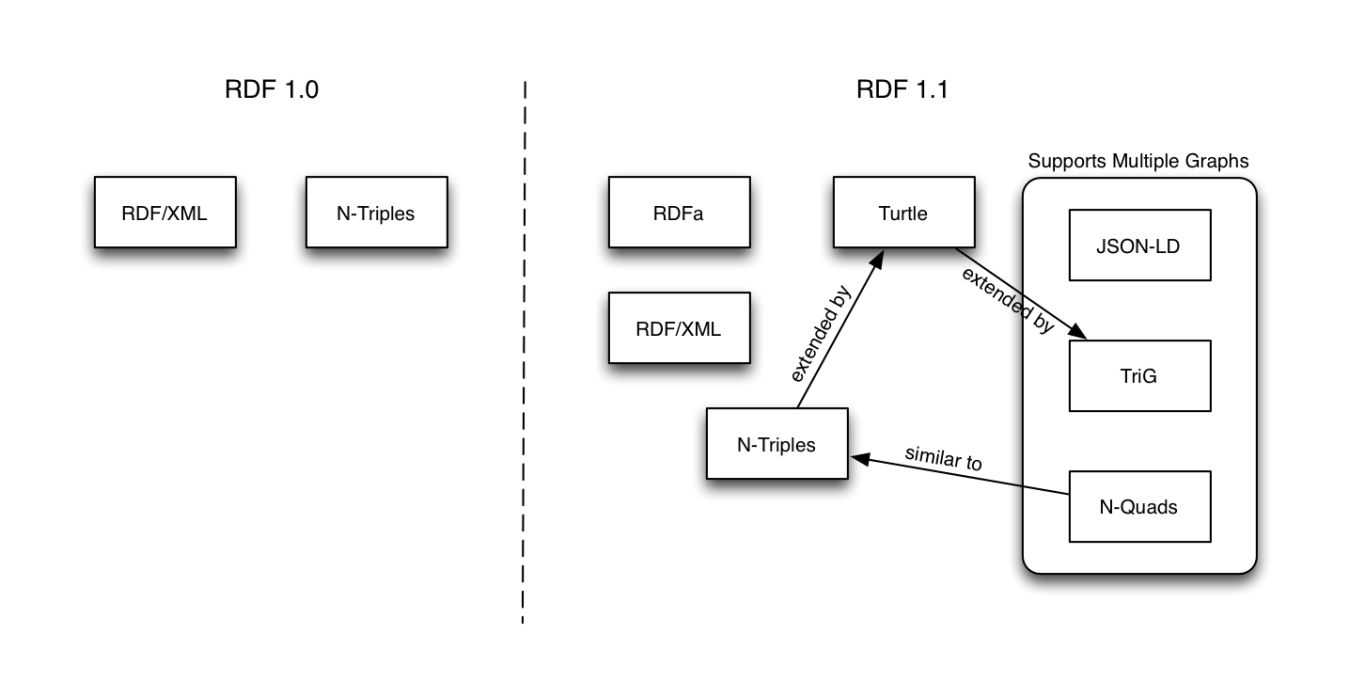


Figura 3. RDF 1.0 y 1.1 formatos de serialización.

Fuente: (Wood, 2014)

* + - * 1. RDF Data Model

La estructura básica de la sintaxis abstracta es un conjunto de triples, cada una consiste de un sujeto, un predicado y un objeto. Un conjunto de triples se llama grafo RDF. Un grafo RDF se puede visualizar como un nodo y el diagrama de arco dirigido, en el que cada uno de triple se representa como un enlace nodo-arco-nodo. (Cyganiak, Wood, & Lanthaler, 2014)

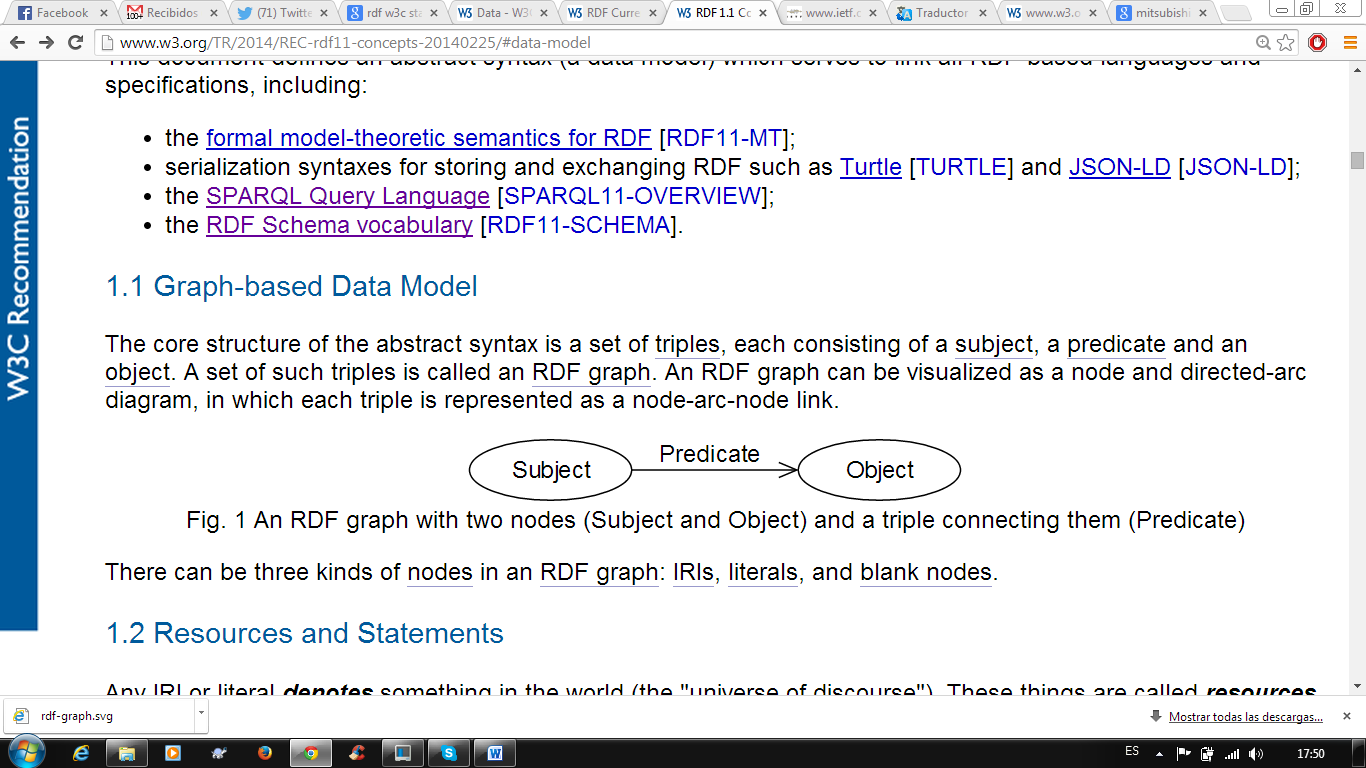


Figura 4: RDF con dos nodos (Subject y Object) y una conexión entre ellos (Predicado).

Fuente: (Cyganiak, Wood, & Lanthaler, 2014)

* + - * 1. RDF Vocabularios y Espacio de Nombres IRIs

Un RDF vocabulario es una colección de IRIs destinados para ser usados en grafos RDF. Por ejemplo, los IRIs documentados en [RDF11-SCHEMA] [[8]](#footnote-8) son el Vocabulario RDF Esquema (RDF Schema Vocabulary). (Cyganiak, Wood, & Lanthaler, 2014)

Una colección de "términos" para un propósito en particular. Los vocabularios pueden ir desde simples el ampliamente usado Schema RDF utilizado, FOAF[[9]](#footnote-9) y Dublin Core Metadata Element Set[[10]](#footnote-10) para vocabularios complejos con miles de términos, tales como los utilizados en la asistencia sanitaria para describir síntomas, enfermedades y tratamientos. Vocabularios juegan un papel muy importante en Linked Data, específicamente para ayudar con la integración de datos. El uso de este término se superpone con la Ontología. (W3C, 2013)

El iris de un vocabulario RDF a menudo comienzan con una subcadena común conocido como un espacio de nombres IRI. Algunos IRIs de espacio de nombres se asocian por convención con un nombre corto conocido como un prefijo de espacio de nombres (namespace prefix). (W3C, 2013)

El término "espacio de nombres" por sí no tiene un significado bien definido en el contexto de la RDF, pero a veces se utiliza de manera informal en el sentido de "espacio de nombres IRI" o "vocabulario RDF”. (W3C, 2013)

Tabla 1. Ejemplos de Prefijos de Espacios de Nombres e IRIs

|  |  |  |
| --- | --- | --- |
| **Namespace prefix** | **Namespace IRI** | **RDF vocabulary** |
| rdf | http://www.w3.org/1999/02/22-rdf-syntax-ns# | The RDF built-in vocabulary [RDF11-SCHEMA][[11]](#footnote-11) |
| rdfs | http://www.w3.org/2000/01/rdf-schema# | The RDF Schema vocabulary [RDF11-SCHEMA][[12]](#footnote-12) |
| xsd | http://www.w3.org/2001/XMLSchema# | The RDF-compatible XSD types[[13]](#footnote-13) |

Fuente: (Cyganiak, Wood, & Lanthaler, 2014)

* + - * 1. Especificación de formatos de serialización de RDF

N-Triples

Tripletas N-Triples son una secuencia de términos RDF que representan al sujeto, predicado y objeto de una Tripleta RDF. Estos pueden estar separados por espacios en blanco (espacios U +0020 o tabulaciones U +0009). Esta secuencia es terminada por un '.' y una nueva línea (opcional al final de un documento). (Carothers & Seaborne, 2014)

N-triple en formato en texto plano para grafos RDF, en la Tabla 2 se ejemplifica su estructura.

Tabla 2. Ejemplo N-triple

|  |  |
| --- | --- |
| 01 | <http://example.org/bob#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> . |
| 02 | <http://example.org/bob#me> <http://xmlns.com/foaf/0.1/knows> <http://example.org/alice#me> . |
| 03 | <http://example.org/bob#me> <http://schema.org/birthDate>  "1990-07-04"^^<http://www.w3.org/2001/XMLSchema#date> . |
| 04 | <http://example.org/bob#me> <http://xmlns.com/foaf/0.1/topic\_interest> <http://www.wikidata.org/entity/Q12418> . |
| 05 | <http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/title> "Mona Lisa" . |
| 06 | <http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/creator> <http://dbpedia.org/resource/Leonardo\_da\_Vinci> . |
| 07 | <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619> <http://purl.org/dc/terms/subject> <http://www.wikidata.org/entity/Q12418>. |

Fuente: (Schreiber & Raimond, 2014)

[Turtle](http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-turtle)

Permite a un grafo RDF a ser completamente escrito en un formulario de texto compacto y natural, con las abreviaturas para los patrones y tipos de datos de uso común. Turtle ofrece niveles de compatibilidad con el formato N-Triples, así como con la sintaxis de patrón de tripletas de la Recomendación de la W3C de SPARQL[[14]](#footnote-14). (Beckett, Berners-Lee, Prud'hommeaux, Carothers, & Machina., 2014)

Turtle es una extensión del N-Triples. Además de la sintaxis básica N-Triples, Turtle introduce una serie de atajos sintácticos, como el soporte para prefijos de espacio de nombres, listas y abreviaturas para datos tipo literales. Turtle ofrece una compensación entre la facilidad de la escritura, la facilidad de análisis y facilidad de lectura. (Schreiber & Raimond, 2014)

Tabla 3. Ejemplo Turtle

|  |  |
| --- | --- |
| 01  02  03  04  05  06  07 | <http://example.org/>  PREFIX foaf: <http://xmlns.com/foaf/0.1/>  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>  PREFIX schema: <http://schema.org/>  PREFIX dcterms: <http://purl.org/dc/terms/>  PREFIX wd: <http://www.wikidata.org/entity/> |
| 08  09  10  11  12  13 | <bob#me>  a foaf:Person ;  foaf:knows <alice#me> ;  schema:birthDate "1990-07-04"^^xsd:date ;  foaf:topic\_interest wd:Q12418 . |
| 14  15  16  17 | wd:Q12418  dcterms:title "Mona Lisa" ;  dcterms:creator <http://dbpedia.org/resource/Leonardo\_da\_Vinci> . |
| 18  19 | <http://data.europeana.eu/item/04802/243FA8618938F4117025F  17A8B813C5F9AA4D619>  dcterms:subject wd:Q12418 . |

Fuente: (Schreiber & Raimond, 2014)

[TriG](http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-trig)

La sintaxis de la Turtle sólo soporta la especificación de grafos simples sin un medio para "nombrarlos". TriG es una extensión de la Turtle que permite la especificación de múltiples grafos en forma de un conjunto de datos RDF. (Schreiber & Raimond, 2014)

Un documento TriG permite escribir un conjunto de datos RDF en una forma textual compacta. Se consiste de una sucesión de directivas, declaraciones triples, declaraciones de grafos que contienen declaraciones triple-generación y líneas en blanco opcionales. (Bizer & Cyganiak, 2014)

Tabla 4. Ejemplo TriG

|  |  |
| --- | --- |
| 01  02  03  04  05  06  07 | BASE <http://example.org/>  PREFIX foaf: <http://xmlns.com/foaf/0.1/>  PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>  PREFIX schema: <http://schema.org/>  PREFIX dcterms: <http://purl.org/dc/terms/>  PREFIX wd: <http://www.wikidata.org/entity/> |
| 08  09  10  11  12  13  14  15  16 | GRAPH <http://example.org/bob>  {  <bob#me>  a foaf:Person ;  foaf:knows <alice#me> ;  schema:birthDate "1990-07-04"^^xsd:date ;  foaf:topic\_interest wd:Q12418 .  } |
| 17  18  19  20  21  22  23  24  25  26 | GRAPH <https://www.wikidata.org/wiki/Special:EntityData/Q12418>  {  wd:Q12418  dcterms:title "Mona Lisa" ;  dcterms:creator <http://dbpedia.org/resource/Leonardo\_da\_Vinci> .    <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619>  dcterms:subject wd:Q12418 .  } |
| 27  28  29 | <http://example.org/bob>  dcterms:publisher <http://example.org> ;  dcterms:rights <http://creativecommons.org/licenses/by/3.0/> . |

Fuente: (Schreiber & Raimond, 2014)

N-Quads

N-Quads es una simple extensión de N-Triples para permitir el intercambio de RDF datasets. N-Quads le permite a uno agregar un cuarto elemento a una línea, capturando en la gráfica IRI la tripleta descrito en esa línea. (Schreiber & Raimond, 2014)

Tabla 5. Ejemplo N-Quads

|  |  |
| --- | --- |
| O1 | <http://example.org/bob#me> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person> <http://example.org/bob> . |
| 02 | <http://example.org/bob#me> <http://xmlns.com/foaf/0.1/knows> <http://example.org/alice#me> <http://example.org/bob> . |
| 03 | <http://example.org/bob#me> <http://schema.org/birthDate> "1990-07-04"^^<http://www.w3.org/2001/XMLSchema#date> <http://example.org/bob> . |
| 04 | <http://example.org/bob#me> <http://xmlns.com/foaf/0.1/topic\_interest> <http://www.wikidata.org/entity/Q12418> <http://example.org/bob> . |
| 05 | <http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/title> "Mona Lisa" <https://www.wikidata.org/wiki/Special:EntityData/Q12418> . |
| 06 | <http://www.wikidata.org/entity/Q12418> <http://purl.org/dc/terms/creator> <http://dbpedia.org/resource/Leonardo\_da\_Vinci> <https://www.wikidata.org/wiki/Special:EntityData/Q12418> . |
| 07 | <http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619> <http://purl.org/dc/terms/subject> <http://www.wikidata.org/entity/Q12418> <https://www.wikidata.org/wiki/Special:EntityData/Q12418> . |
| 08 | <http://example.org/bob> <http://purl.org/dc/terms/publisher> <http://example.org> . |
| 09 | <http://example.org/bob> <http://purl.org/dc/terms/rights> <http://creativecommons.org/licenses/by/3.0/> . |

Fuente: (Schreiber & Raimond, 2014)

JSON-LD (JSON-based RDF syntax);

Proporciona una sintaxis JSON para gráficos y conjuntos de datos RDF. JSON-LD puede ser utilizado para transformar documentos JSON a RDF con cambios mínimos. JSON-LD ofrece identificadores universales de objetos JSON, un mecanismo en el que un documento JSON se puede referir a un objeto descrito en otro documento JSON en otros lugares en la Web, así como el tipo de datos y el lenguaje de manipulación.

Tabla 6. Ejemplo JSON-LD

|  |  |
| --- | --- |
| O1 | { |
| 02 | "@context": "example-context.json", |
| 03 | "@id": "http://example.org/bob#me", |
| 04 | "@type": "Person", |
| 05 | "birthdate": "1990-07-04", |
| 06 | "knows": "http://example.org/alice#me", |
| 07 | "interest": { |
| 08 | "@id": "http://www.wikidata.org/entity/Q12418", |
| 09 | "title": "Mona Lisa", |
| 10 | "subject\_of": "http://data.europeana.eu/item/04802/ 243FA8618938F4117025F17A8B813C5F9AA4D619", |
| 11 | "creator": "http://dbpedia.org/resource/Leonardo\_da\_Vinci" |
| 12 | } |
| 13 | } |

Fuente: (Schreiber & Raimond, 2014)

[RDFa](http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-rdfa)

Es una sintaxis de RDF que se puede utilizar para insertar datos RDF dentro de los documentos HTML y XML. Esto permite, por ejemplo, a los motores de búsqueda agregar estos datos al rastrear la Web y enriquecer los resultados de búsqueda. (Schreiber & Raimond, 2014)

Tabla 7. Ejemplo RDFa

|  |  |
| --- | --- |
| O1 | <body prefix="foaf: http://xmlns.com/foaf/0.1/ |
| 02 | schema: http://schema.org/ |
| 03 | dcterms: http://purl.org/dc/terms/"> |
| 04 | <div resource="http://example.org/bob#me" typeof="foaf:Person"> |
| 05 | <p> |
| 06 | Bob knows <a property="foaf:knows" href="http://example.org/alice#me">Alice</a> |
| 07 | and was born on the <time property="schema:birthDate" datatype="xsd:date">1990-07-04</time>. |
| 08 | </p> |
| 09 | <p> |
| 10 | Bob is interested in <span property="foaf:topic\_interest" |
| 11 | resource="http://www.wikidata.org/entity/Q12418">the Mona Lisa</span>. |
| 12 | </p> |
| 13 | </div> |
| 14 | <div resource="http://www.wikidata.org/entity/Q12418"> |
| 15 | <p> |
| 16 | The <span property="dcterms:title">Mona Lisa</span> was painted by |
| 17 | <a property="dcterms:creator" href="http://dbpedia.org/resource/Leonardo\_da\_Vinci">Leonardo da Vinci</a> |
| 18 | and is the subject of the video |
| 19 | <a href="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619">'La Joconde à Washington'</a>. |
| 20 | </p> |
| 21 | </div> |
| 22 | <div resource="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619"> |
| 23 | <link property="dcterms:subject" href="http://www.wikidata.org/entity/Q12418"/> |
| 24 | </div> |
| 25 | </body> |

Fuente: (Schreiber & Raimond, 2014)

[RDF/XML](http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/#section-rdfxml)

Proporciona una sintaxis XML para grafos RDF. Cuando RDF ​​fue desarrollado originalmente en la década de 1990, esta fue su única sintaxis, y algunas personas siguen llamando esta sintaxis "RDF". En 2001, se propuso un precursor de la Tortuga llamado "N3", y poco a poco los otros idiomas que aparecen aquí se han adoptado y normalizado. (Schreiber & Raimond, 2014)

Tabla 8. Ejemplo RDF/XML

|  |  |
| --- | --- |
| O1 | <?xml version="1.0" encoding="utf-8"?> |
| 02 | <rdf:RDF |
| 03 | xmlns:dcterms="http://purl.org/dc/terms/" |
| 04 | xmlns:foaf="http://xmlns.com/foaf/0.1/" |
| 05 | xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" |
| 06 | xmlns:schema="http://schema.org/"> |
| 07 | <rdf:Description rdf:about="http://example.org/bob#me"> |
| 08 | <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/> |
| 09 | <schema:birthDate rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1990-07-04</schema:birthDate> |
| 10 | <foaf:knows rdf:resource="http://example.org/alice#me"/> |
| 11 | <foaf:topic\_interest rdf:resource="http://www.wikidata.org/entity/Q12418"/> |
| 12 | </rdf:Description> |
| 13 | <rdf:Description rdf:about="http://www.wikidata.org/entity/Q12418"> |
| 14 | <dcterms:title>Mona Lisa</dcterms:title> |
| 15 | <dcterms:creator rdf:resource="http://dbpedia.org/resource/Leonardo\_da\_Vinci"/> |
| 16 | </rdf:Description> |
| 17 | <rdf:Description rdf:about="http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619"> |
| 18 | <dcterms:subject rdf:resource="http://www.wikidata.org/entity/Q12418"/> |
| 19 | </rdf:Description> |
| 20 | </rdf:RDF> |

Fuente: (Schreiber & Raimond, 2014)

* + - 1. SPARQL Query Language for RDF

SPARQL se puede utilizar para expresar consultas que permiten interrogar diversas fuentes de datos, si los datos se almacenan de forma nativa como RDF ​​o son definidos mediante vistas RDF ​​a través de algún sistema middleware. SPARQL contiene las capacidades para la consulta de los patrones obligatorios y opcionales de grafo, junto con sus conjunciones y disyunciones. SPARQL también soporta la ampliación o restricciones del ámbito de las consultas indicando los grafos sobre los que se opera. Los resultados de las consultas SPARQL pueden ser conjuntos de resultados o grafos RDF. (Prud'hommeaux & Seaborne, 2008)

La mayoría de las formas de consulta en SPARQL contienen un conjunto de patrones de tripleta (triple patterns) denominadas patrón de grafo básico. Los patrones de tripleta son similares a las tripletas RDF, excepto que cada sujeto, predicado y objeto puede ser una variable. Un patrón de grafo básico concuerda con un subgrafo de datos RDF cuando los términos RDF (RDF terms) de dicho subgrafo pueden ser sustituidos por las variables y el resultado es un grafo RDF equivalente al subgrafo en cuestión. (Prud'hommeaux & Seaborne, 2008)

A continuación se redacta un ejemplo de una consulta SPARQL, tomado de (Prud'hommeaux & Seaborne, 2008).

Datos:

<http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> "SPARQL Tutorial" .

Consulta:

SELECT ?title

WHERE

{

<http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> ?title.

}

Resultado de la consulta:

Tabla 9. Resultado consulta SPARQL

|  |
| --- |
| **title** |
| "SPARQL Tutorial" |

Fuente: (Prud'hommeaux & Seaborne, 2008).

* + 1. Acerca de DBpedia

La DBpedia[[15]](#footnote-15) nos da la siguiente definición sobre si misma: “Es un esfuerzo de la comunidad crowd-sourced[[16]](#footnote-16) para extraer información estructurada de Wikipedia[[17]](#footnote-17) y hacer esta información disponible en la web. DBpedia permite que hacer consultas sofisticadas contra Wikipedia.”

El conocimiento extraído de Wikipedia es publicado cumpliendo los estándares de la Web Semántica y las mejores prácticas de Linkend Data.

* + - 1. Framework extracción

Los artículos de Wikipedia consisten sobre todo en texto libre, pero también comprenden diversos tipos de información estructurada en forma de wiki markup[[18]](#footnote-18). Dicha información incluye plantillas infobox, información de categorización, imágenes geo-coordenadas, enlaces a páginas web externas, páginas de desambiguación, redirecciones entre páginas y vínculos a través de las diferentes ediciones lingüísticas de Wikipedia. El marco de la extracción DBpedia extrae esta información estructurada de Wikipedia y lo convierte en una rica base de conocimientos (Lehmann, y otros, 2012)

En la figura 2 se observa el marco de trabajo necesario para lograr que todo el proceso partiendo de la extracción de información desde Wikipedia hasta poder disponer de ella como datos enlazados.

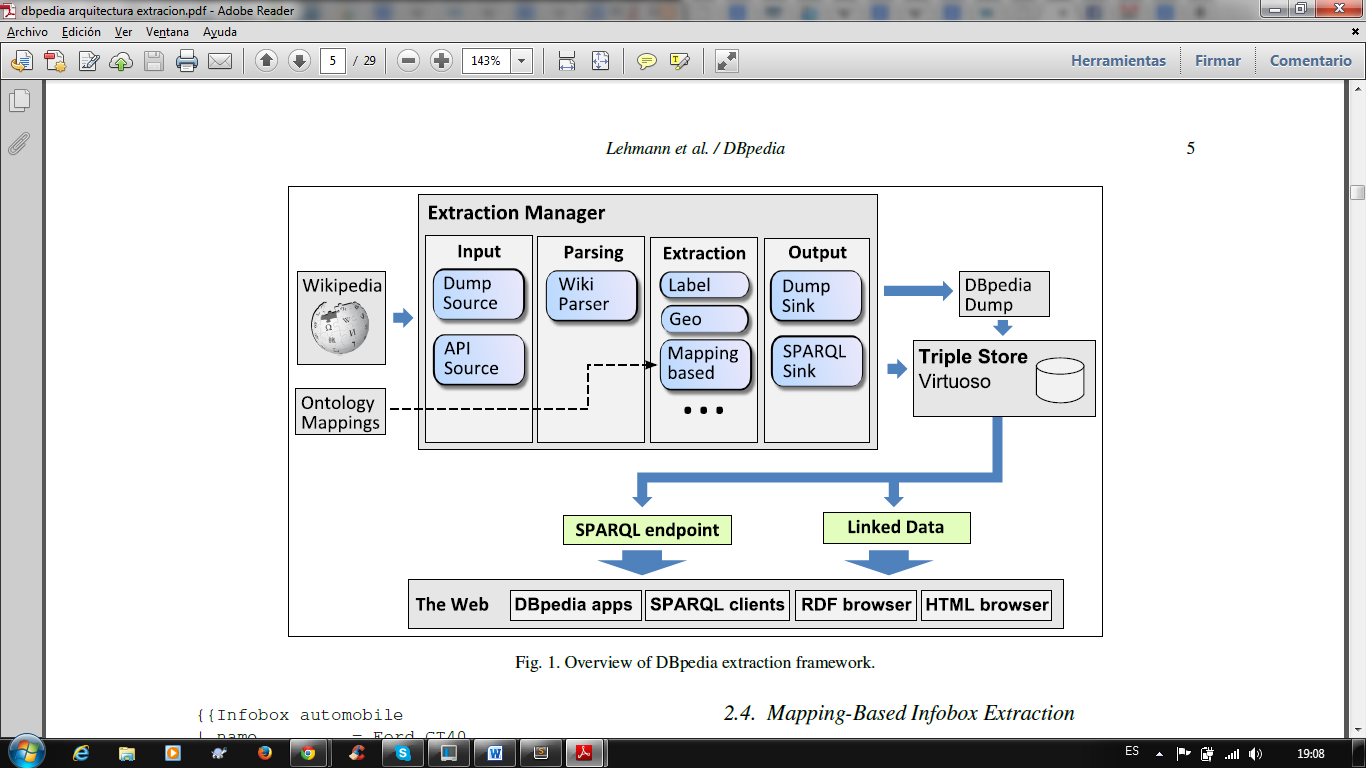


Figura 5: Representación gráfica RDF

Fuente: (Lehmann, y otros, 2012)

* + - 1. DBpedia Dataset

DBpedia se trata de una base de conocimiento (en inglés knowledge base) que se encuentra distribuida 119 idiomas que en total describen 12.6 millones de cosas únicas, 24.6 millones de enlaces a imágenes 27.6 millones enlaces a fuentes externas 45 millo es de enlaces a fuentes externas de datos RDF y 67 millones de enlaces a categorías Wikipedia 42.1 millones a categorías YAGO[[19]](#footnote-19).

Lo cual la establece como una fuente muy basta de información sobre cualquier ámbito de conocimiento, esto gracias a al continuo crecimiento de la Wikipedia, su fuente de información. Pero no esto no quiere decir que única base de conocimiento disponible en la web, podemos encontrar otras como YAGO.

* + - 1. Acceso a DBpedia Dataset

El dataset de Dbpedia se almacena y publica mediante OpenLink Virtuoso. La infraestructura de Virtuoso permite el acceso a los datos RDF de DBpedia a través de SPARQL endpoint, junto al soporte HTTP para cualquier GET estándar de cliente Web para HTML o representación RDF de un recurso Dbpedia. (Bizer, Dbpedia, 2009).

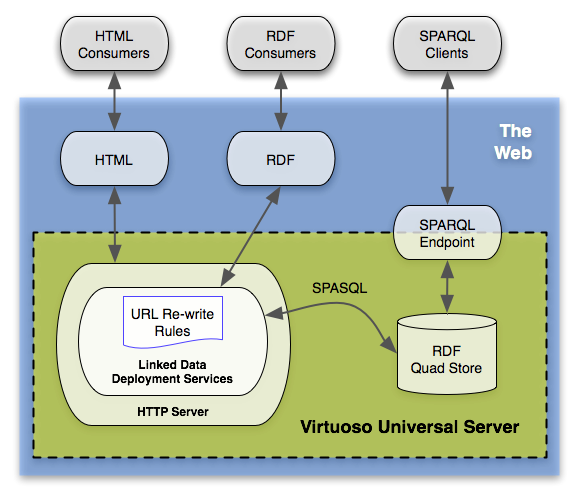


Figura 6. Arquitectura de provisión de Datos de Dbpedai.

Fuente: (Bizer, Dbpedia, 2009)

Se expones dos formas para acceder a la a la dataset de dbpedia:

1. Querying DBpedia: se puede accederé a través del SPARQL endpoint público http://dbpedia.org/sparql proporcionado por OpenLink Virtuoso. Por este método podemos acceder enviando query Sqparl para hacer consultar sobre dataset.
2. Linked Data: se refiere a la aplicación de los principios de datos enlazados revisados en 1.1.2. para nombrar y referenciar los recursos dentro de dataset de dbpedia como por ejemplo: http://dbpedia.org/resource/The\_Lord\_of\_the\_Rings

Además de esta opciones podemos descargar el dataset de dbpedia en diferentes idiomas teniendo en cuenta de que el número de recursos puede cambiar de idioma a idioma puesto que no se trata de una traducción sino de una recopilación de información de Wikipedia la cual se encuentra más extendida en ingles que otros lenguas,

* 1. Procesamiento de Lenguaje Natural (PLN)
     1. Introducción

El procesamiento de lenguaje natural dentro de la ingeniería lingüística comprende la rama se preocupara por entender el lenguaje humano, una tarea que para las personas e inclusive animales es tan natural y cotidiana se vuelve un reto al tratar de interpretarlo mediante procesos computacionales a fin de comprenderlo y poder replicarlo.

La dificultada de la construcción de una aplicación de la ingeniería lingüística variara de acuerdo a objetivo que se persiga, esto explicado por (García, 2005) en donde ejemplifica: “un sistema de generación de cartas personalizado no precisa ningún tratamiento de comprensión, o un sistema de identificación de la lengua (o un detector de errores ortográficos) no necesitan generar lenguaje humano. La mayoría de las aplicaciones incluyen, sin embargo, alguna forma más o menos precisa de comprensión. Así, un sistema de consulta en lenguaje humano a una base de datos precisa un nivel muy alto de comprensión de las expresiones del interlocutor humano para que la respuesta del sistema sea de utilidad. En cambio, en un sistema de traducción o de resumen automáticos se pueden lograr niveles de corrección muy notables con niveles de comprensión bajos. Es decir, no es preciso comprender totalmente una oración para ser capaz de traducirla correctamente.”

* + 1. Part of Speech Tagger

También denominado POS-tagging Nos permite distinguir la función de una palabra en un determinada contexto mediante la asignación de una etiqueta predefinida. (Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P., 1992) Nos dicen que: “Una part of speech tagger es un sistema que usa el contexto para asignar parte de un discurso a una palabra”.

Es te etiquetado de palabra ya permite un primera desambiguación en cuanto a la función de la palabra en un sentencia o contexto. Asi podemos por ejemplo[[20]](#footnote-20) ver que la palabra “*dado”* que si bien es nombre en singular, también puede ser una foram del verbo dar.

Pero antes de poder etiquetar una palabra por su función es necesaria una tokenización del texto que va a analizar, que consiste en separarlo en palabras individuales reconociendo un token para palabra o carácter extraído.

* + 1. Chunking

Text Chunking consiste en dividir un texto en frases de tal manera que palabras sintácticamente relacionadas sean miembros de la misma clase. Estas frases no se superponen es decir que una sola palabra puede ser miembro de un chunk. (Tjong Kim Sang, E. F., & Buchholz, S., 2000)

Este proceso es básico al momento de detectar entidades dentro de un texto, este proceso lo podemos observar en la figura 1 en donde la sentencia, *We saw the yellow dog*, está separada en cuadros en los más pequeños observamos etiquetas de POS Tag y las más grades al nivel de chunking.Una vez la frase ha pasado por el proceso de chunking podemos rescatar dos entidades dentro de la sentencia como *We* y *the yellow dog.*

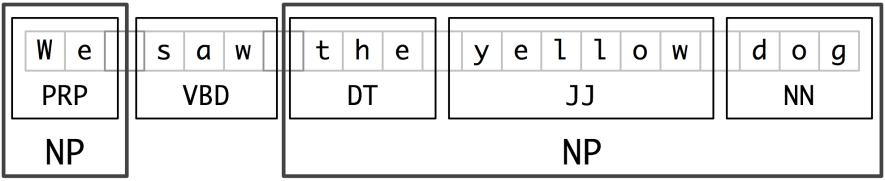


Figura 7: Ejemplo POS Tag y Chunking

Fuente: http://www.nltk.org/book/ch07.html

* + 1. Desambiguación

Este es un fenómeno muy común conocido polisemia que se refiere a cuando una palabra tiene varios significados la desambiguación busca descifrar que significado es el que está activado en determinado contexto se denomina Word Sense Disambiguation (WSD), este problema propio del procesamiento de lenguaje natural (PLN). El descifrar estos distintos significados para los seres humanos es muy común, lo resolvemos de forma cotidiana y pasa casi desapercibida.

* + - 1. Métodos basados en el conociendo.

El método Lesk (Lesk 1986) es el método basado en el diccionario. Se basa en la hipótesis de que las palabras usadas juntas en el texto están relacionadas entre sí y que la relación se puede observar en las definiciones de las palabras y sus sentidos. Dos (o más) palabras son desambiguadas encontrando el par de sentidos diccionario con la palabra mayor superposición en sus definiciones del diccionario.

Diccionarios externos:

*WordNet:*

Es una gran base de datos léxica de inglés. Sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos (synsets), cada una expresando un concepto distinto. Synsets están vinculados entre sí por medio de las relaciones conceptuales semántico y léxico.

*Corpus:*

Corpus es una gran colección de textos. Se trata de un cuerpo de material escrito o hablado sobre la que se basa un análisis lingüístico. Es un conjunto muy amplio de ejemplos reales del uso de la lengua.

*WordNet Domains:*

WordNet Domains es una extensión de WordNet 1.6, donde cada synset tiene asociado uno o varios dominios (categorías semánticas). Estos dominios, están clasificados en una jerarquía con distintos niveles de especialización, cuanto más profundo es el nivel sobre el que nos movemos, mayor es el grado de especialización.

* + - 1. Métodos supervisados

Métodos supervisados ​​se basan en la suposición de que el contexto puede proporcionar evidencia suficiente por sí sola para eliminar la ambigüedad de las palabras. Probablemente cada algoritmo de aprendizaje automático va se ha aplicado a WSD, incluyendo técnicas asociadas tales como la selección de características, la optimización de parámetros, y el aprendizaje conjunto.

* + - 1. Métodos semi-supervisados

El enfoque de bootstrapping comienza a partir de una pequeña cantidad de datos de semillas para cada palabra: cualquiera de ejemplos de entrenamiento manualmente etiquetados - o un pequeño número de reglas de decisión de éxito seguro. Las semillas se utilizan para entrenar un clasificador inicial, utilizando cualquier método supervisado. Este clasificador se utiliza en la parte sin etiqueta del corpus para extraer un conjunto de entrenamiento más grande, en el que sólo se incluyen las clasificaciones más seguras. El proceso se repite, cada nuevo clasificador siendo entrenado en un corpus de entrenamiento sucesivamente mayores, hasta que se consume todo el corpus, o hasta que se alcanza un número máximo dado de iteraciones.

* + - 1. Métodos sin supervisión

Aprendizaje no supervisado es el mayor desafío para los investigadores WSD. El supuesto subyacente es que los sentidos similares ocurren en contextos similares, y por lo tanto los sentidos puede ser inducido a partir del texto agrupando las ocurrencias de palabras usando alguna medida de similitud de contexto.

* 1. Servicios Web
     1. Introducción

La W3C[[21]](#footnote-21) (World Wide Web Consortium) encarda de estandarización de las tecnologías en la web aborda este tema de la siguiente forma: “Los servicios web proporcionan un medio estándar de interoperabilidad entre las distintas aplicaciones de software, que se ejecuta en una variedad de plataformas y/o marcos de trabajo. Los servicios Web se caracterizan por su gran interoperabilidad y extensibilidad. Se pueden combinar en una forma de acoplamiento flexible con el fin de lograr operaciones complejas. Programas que prestan servicios simples pueden interactuar entre sí con el fin de ofrecer servicios de valor añadido sofisticados.” Los servicios web permiten la colaboración entre aplicaciones independientemente de la plataforma en las que están desarrolladas, utiliza protocolos y normas estandarizadas en la web, además esto permite la reutilización de código, además de disminuir el coste de integración.

* + 1. Tipos de Web Services

Encontramos dos tipos de servicios web de acuerdo con la forma en que se puede implementar abarcado diferentes tecnologías: *RESTful Web Services* y *“Big”[[22]](#footnote-22) Web Services* (o también, The “Big” Web services technology stack, debido a la diversas tecnologías en las que se implementar como: SOAP, WSDL, WS-Addressing, WS-ReliableMessaging, WSSecurity, etc), estos dos tipos son expuesto y comparados en (Pautasso, Zimmermann, & Leymann, 2008)

* + - 1. SOAP AND THE WS-\* STACK

Proporcionar interoperabilidad sin fisuras entre los heterogéneos pilas de tecnología de middleware y el fomento de la articulación flexible de servicio al consumidor (solicitante, cliente) y proveedor de servicios son los principales objetivos de diseño de arquitectura orientada a servicios (SOA) conceptos y tecnologías de servicios Web. (Pautasso, Zimmermann, & Leymann, 2008)

En el plano conceptual, un servicio es un componente de software que se proporciona a través de un endpoint[[23]](#footnote-23) accesible en la red. Consumidores de servicios y proveedores usan mensajes para intercambiar solicitudes e información de respuesta en forma de *documentos self-containing[[24]](#footnote-24)* que hacen muy pocas suposiciones sobre las capacidades tecnológicas del receptor. En particular, no hay noción de una referencia de objeto remoto que requeriría un corredor de objeto para gestionar un espacio distribuido dirección de memoria. En el nivel de la tecnología, SOAP es un lenguaje XML que define una arquitectura de mensajes y formatos de mensaje, por lo tanto, proporcionar un protocolo de procesamiento rudimentario. El documento SOAP define un elemento XML de nivel superior llamada sobre, que contiene un encabezado y un cuerpo. El encabezado SOAP es un contenedor de información de infraestructura extensible de capa de mensajes que se puede utilizar para fines de enrutamiento (por ejemplo, hacer frente) y Calidad de Servicio (QoS) de configuración (por ejemplo, las transacciones, la seguridad, la fiabilidad). El cuerpo contiene la carga útil del mensaje. Esquema XML se usa para describir la estructura del mensaje SOAP, por lo que los motores de jabón en los dos puntos finales pueden Marshall y Resolver referencia el contenido del mensaje y la ruta a la aplicación apropiada. (Pautasso, Zimmermann, & Leymann, 2008)

* + - 1. REST

Transferencia de estado representacional (REST) ​​se introdujo originalmente como un estilo de arquitectura para la construcción de sistemas hipermedia distribuidos a gran escala. Este estilo arquitectónico es una entidad más abstracta, cuyos principios se han utilizado para explicar la excelente escalabilidad del protocolo HTTP 1.0 y también han limitado el diseño de su siguiente versión, HTTP 1.1. Por lo tanto, el término REST muy a menudo se utiliza junto con HTTP. (Pautasso, Zimmermann, & Leymann, 2008)

REST no es un protocolo, un formato de archivo, o un marco de desarrollo. Es un conjunto de restricciones de diseño, la hipermedia como el motor de estado de la aplicación.

El estilo arquitectónico REST se basa en cuatro principios:

*Identificación de recursos a través de URI*. Un servicio web RESTful expone un conjunto de recursos que identifican los objetivos de la interacción con sus clientes. Los recursos son identificados por URI, que proporcionan un espacio de direccionamiento global de los recursos y de descubrimiento de servicios.

*Interfaz uniforme*. Los recursos son manipulados utilizando un conjunto fijo de cuatro crear, leer, actualizar, eliminar operaciones: PUT, GET, POST y DELETE. PUT crea un nuevo recurso, que puede ser luego borrar con DELETE. GET recupera el estado actual de un recurso en alguna representación. POSTE transfiere un nuevo estado sobre un recurso.

*Mensajes de auto-descriptivo*. Recursos están desconectados de su representación para que su contenido se puede acceder en una variedad de formatos (por ejemplo, HTML, XML, texto plano, PDF, JPEG, etc.) Metadatos sobre el recurso está disponible y se utiliza, por ejemplo, para controlar el almacenamiento en caché, detectar errores de transmisión, negociar el formato de representación adecuada, y llevar a cabo la autenticación o controlar el acceso. Interacciones con estado a través de hipervínculos. Cada interacción con un recurso no tiene estado, es decir, los mensajes de solicitud son autónomos.

Interacciones con estado se basan en el concepto de transferencia de estado explícito. Existen varias técnicas para el intercambio de estado, por ejemplo, la reescritura de URI, cookies, y los campos de formulario ocultos. Estado puede ser embebido en los mensajes de respuesta para señalar válidos estados futuros de la interacción.

* + - 1. Hipermedia

(Ruby, 2007)Es una estrategia que nos permite establecer una conexión entre los recursos y describe sus capacidades, La estrategia hipermedia tiene siempre el mismo objetivo. Hipermedia es una manera para que el servidor para decirle al cliente qué HTTP request el cliente podría querer hacer en el futuro. Es un menú, proporcionado por el servidor, desde el que el cliente es libre de elegir. El servidor sabe lo que podría pasar, pero el cliente decide lo que realmente sucede.

* + - 1. Recursos y representaciones

(Leonard Richardson and Mike Amundsen, 2013)Rest denomina recursos a los datos estructurados que son objetos de las interacciones entre métodos de HTTP, y se dice que cualquier cosa que pueda ser almacenado de un computador puede ser un recurso, como documento electrónico, una fila de una base de datos o el resultado de un algoritmo

No solo las cosas almacenadas en un computador pueden ser llamados recursos también pueden ser recursos artículos tangibles como frutas por ejemplos, y es posible representarlo como recursos a través de la web como por ejemplo como un artículo en de venta o una imagen binaria depende de la aplicación así que por eso decimos sobre las representaciones que puede ser cualquier documento legible que contenga información acerca de un recurso.

* + - 1. URI y relación con los recursos

(Leonard Richardson and Mike Amundsen, 2013)Este tecnología de propia de la web que ya hemos revisa en capítulos anterior y que ahora recordamos teniendo en cuenta de que Rest trabaja sobre recursos o representaciones de los mismos u que la forma estandarizada para la identificación los recursos son las URIs, así como la relación existencial que tiene una sobre la otra.

* 1. Trabajos relacionados

# CAPITULO 2: PROBLEMÁTICA

* 1. Estado actual

La documentación dentro del desarrollo de trabajos educativos universitarios es indispensable para la difusión de los avances realizados en las distintas ramas y disciplinas de las ciencias, estas publicaciones se encuentran en texto plano el cual está diseñado para la fácil compresión por parte del usuario humano que acceda a estos, y contienen datos relevantes dentro de sus líneas los cuales se pierden puesto que no son explotados. Y que a pesar que estos documentos de encontrasen almacenados y publicados de tal forma que sean alcanzados por motores de búsquedas que se especializan en el procesamiento de texto, los datos relevantes que no son accesibles y que representa recursos disponibles en la Web a los cuales se quiere enlazar.

Se busca que las fuentes de datos que contengan estos recursos “objetivos”, sean estructurados y publicados de acuerdo con los principios de Datos Enlazados y permita acceder a Linking Open Data cloud.

En la estructura de estos documentos existe resumen inicial en el que se explica el tema que se abarca en la publicación, etiquetado en inglés como “abstract” que en español significa “resumen” que es donde se centraran los esfuerzos para descubrir datos, esto a pesar de la existencia otros campos como “keywords” (en español, palabras claves), que exponen los temas que expone los abordan, pero que no son tan descriptivos.

* 1. Justificación

Teniendo en cuanta la tendencia actual de web, la web semántica que se basa en los principios de datos enlazados, los datos toma un factor importante, por lo cual que estos se encuentren “ocultos” dentro del texto no hace posible su enlace e impide la apertura hacia otras fuentes de información.

Esto hace necesario medios que permitan extraer y relacionar estos datos dentro de las publicaciones, de acuerdo a los principio de la web semántica que se encuentra en construcción y que iniciativas como esta ayudan a su expansión.

La información escrita es de fácil compresión para las seres humanos, se puede entender palabras por palabra su significado, aun cuando este puede varias de acurdo al contexto en que se encuentre y a la vez modificando en significado de otras palabras.

Dentro de un texto existen palabras que son más representativas que otras al momento de dar sentido a toda una sentencia o frase, esto puede ocurrir debido a que una palabra o varias palabras, más halla de tener un sentido pueden ser representaciones que entidades existentes en mundo real, como: personas, lugares, eventos, organizaciones etc. o representen entidades abstractas como la Web y diferentes tecnologías existentes, en sí, un texto plano como tal pude estar relacionado con diferentes entidades del mundo real por pedio de las palabras.

Pero el sentido de una frase descansa todas las palabras siendo unas más representativas que otras como ya habíamos visto, aunque no necesariamente estas tengas representaciones, sino los diferentes temas que abarca el texto.

* 1. Objetivos
* Desarrollo de Servicio Web REST
  + Desambiguación y Enlace (Descubrimiento y Enriquecimiento Datos LOD-Cloud)
* Construcción de una base de conocimientos
* Diseño y Desarrollo de Frontal Piloto para Integración de Servicios Web

CAPITULO 3: Solución

1. 1. Propuesta

Los datos que se encuentran dentro del texto tanto de las publicaciones como de fuentes en general, se encurtan relacionados con otros temas y fuentes de datos, a los que por medio este proyecto se tratara de acceder, extraer, relacionar y enlazar con fuentes de información abierta como lo es DBpedia (que se basa en los principios de los Datos Enlazados), esto permitirá el enriquecimiento del contenido. A todos estos datos a descubrir los denominaremos

Esta propuesta utiliza Servicios Web Rest lo que permite una independencia de la fuente de origen de texto a ser analizado, con la lógica de la aplicación propuesta y esta a su vez devuelve un resultado en que en este caso será formato Json[[25]](#footnote-25) (JavaScript Object Notation - Notación de Objetos de JavaScript) que es ampliamente conocido y utilizado para el intercambio de datos, hacia el cliente que consume el servicio.



Figura 8. Lógica propuesta para la Aplicación;

Fuente: (Propio)

La lógica que se propone es explicada en la Figura 8, en donde las interacciones inician con el ingreso del texto a ser analizado, en el cual se aplican las diferentes tecnologías revisadas en Capítulos 1 de este documento, para obtener como resultado entidades y keywords estructurados en formato JSON, que estarán desambiguadas y enlazadas, de existir un recurso al cual corresponda dentro del DataSet de DBpedia, es decir que no todos las entidades que encontremos dentro del texto de una publicación es referenciado en la DBpedia.

Puesto que nos concentraremos en los *abstracts* de las publicaciones y que esto se redactados en idioma ingles a pesar de que se trate de una publicación en español, se limitara el desarrollo de la soluciona este idioma.

Esta propuesta se surge como solución para el descubrimiento de datos en el texto de los *abstracts* de las publicaciones universitarias, pero debido a la gran cantidad de recursos de diversos temas que se encuentran actualmente disponibles en DBpedia, se puede aplicar a cualquier texto (en idioma ingles) para esto se implementa un interfaz gráfica web donde el usuario puede insertar su texto y ver el resultado.

* 1. Arquitectura

La arquitectura se implementa en capas, haciendo un división de los componentes que en la lógica de la aplicación interactúan para dar solución a la problemática, estos los podemos encontrar en el Figura 9. Las capas que componen la arquitectura de aplicación se detallan a continuación.



Figura 9. Arquitectura.

Fuente: (Propio)

* + 1. Lógica de negocios (Aplicación)

Cumple con implementación de la lógica con la resuelve la problemática, encargado de resolver las solicitudes por parte de los clientes que se encuentran del lado de vistas de la aplicación.

Desarrollado en Python versión 2.7.4

* + 1. Dataset Local con recursos de DBpedia y base de conocimientos

El poder obtener accesos a los recursos de la DBpedia es vital para dar desarrollar la propuesta, y debido a que la conexión con la DBpedia es un obstáculo puesto que estas no siempre está disponible o su retorno de resultados son lentos, se decidió implementar un servidor local propio que contenga, los recursos y atributos de estos, necesarios para poder ser encontrados, desambiguados y referenciados.

DBpedia utiliza Virtutoso OpenLink para almacenar y servir los datos que contiene, por lo cual se decidio utilizar esta misma tecnología para el servidor local utilizando su versión Open Source, como almacén de datos y poder replicar el comportamiento de la misma en la web.

Archivos necesarios para implementar los recursos:

* Labels de recursos:
  + labels\_en.nt.bz2
* Datos personales de los recursos tipos Persona:
  + persondata\_en.nt.bz2
* Resúmenes Corto de los recursos
  + short\_abstracts\_en.nt.bz2
* Links de Desambiguación de Wikipedia
  + disambiguations\_en.nt.bz2
* Redirecciones entre Recursos
  + redirects\_en.nt.bz2

Debido a que cada archivo una vez descomprimido excede la capacidad computacional para la importación hacia Virtuoso OpenLink en el servidor local se procedió a dividir los archivos de importación en un tamaño no mayor a 20 MB, el tamaño total de las importaciones es de 3.8 GB, de espacio en disco.

* + 1. Frontal

La vista es un Interfaz Web que permite al usuario la facilidad de la comunicación entre la aplicación y el usuario. Instruye al usuario sobre el uso de la herramienta, identificando con facilidad los parámetros necesarios y en especial la facilidad de la presentación de los resultados. Permite la integración de los compontes del sistema en un entorno amigable para el usuario.

Prototipo de interfaz

* 1. Metodología
  2. Riesgos
  3. Módulos
     1. Validación de Texto

El proceso de validad de texto permite fundamentalmente evitar errores posteriores, verificando que el parámetro de texto de entrada se encuentre bien estructurado.

* + 1. Módulo de etiquetado

Una vez que el texto como parámetro ha sido validado ya es posible trabajar con él, este módulo es el primero encargado de procesarlo a través PLN se realizan dos actividades importantes como es el etiquetado como tal y un paso antes de eso la tokenización del texto en sentencias cortas, si este es un párrafo y consta de más de una oración, y luego en palabras validas separando palabra por palabra, *tokens*.

**Diagramo de secuencia**



Figura 10. Diagrama de secuencias del módulo de Etiquetado

Fuente: (Propio)

A continuación se detalla los componentes del módulo.

**Main**: en proceso que inicio a las interacciones entre los componentes del sistema, que será una solicitud por parte de un usuario haciendo uso de un cliente REST, al cual se devolverán los resultados de los diferentes módulos del sistema. Esta función se encuentra presente en todos los módulos que se verán a continuación puesto que es la que gestiona la comunicación entre estos.

**Procesar Texto**: valida y prepara el parámetro texto para su procesamiento, descrito en 3.5.1, luego de ser validado su resultado es devuelto al proceso principal *Main*.

**Tokenizar**: permite de división del texto en sentencias corta para analizar por separado, y luego en palabras o tokens para procesar al etiquetado.

**Etiquetar**: toma como entrada el texto tokenizado, para proceder a realizar el etiquetado Part-of-Speech, y así asignar mediante una etiquetas la función de cada palabra dentro del contexto en el que se encuentra y este resultado devuelve al proceso principal.

* + 1. Extracción de Entidades y Keywords

Permite extraer mediante Procesamiento de Lenguaje Natural (PNL) las entidades de quienes se trata en texto y las principales palabra que las acompañan.

**Diagrama de secuencias**



Figura 11. Diagrama de secuencias del módulo de Extracción de Entidades y Keywords

Fuente: (Propio)

Este módulo tiene como entrada el parámetro de salido del módulo de Etiquetado, continuando con el procesamiento del texto, el detalle de los elementos de este módulo se describen a continuación.

**Extracción Entidades:** permite reconocer los elementos importantes dentro de la redacción del texto y extraerlos esto en base al etiquetado realizado en 3.5.2, estos elementos y el texto pasan como parámetro para funciones siguientes.

**Extracción Keyworks Simples:** permite extraer palabreas importantes del contenido de forma individual aunque en el texto se encuentren acompañas por otras palabras, su resultado asi como el texto son para metro de estrada para la siguiente proceso.

**Extracción Keyworks Compuestas:** identifica y extrae palabras compuestas por más de una palabra, el resultado es devuelto hasta la función junto con las entidades y las palabras importantes individuales Main.

* + 1. Desambiguación y Enlace

Este módulo pensado en un principio en dos desambiguación por un lado y enlace por otro, se unificado debido a adaptación que se realiza al algoritmo de **Lesk** (descrito en *1.2.4.1* métodos de desambiguación basados en conocimiento) donde los recursos y las descripciones de las mismos en DBpedia, son tomados como base del conocimiento, para determinar el sentido y correspondencia (con un recurso de DBpedia) de una entidad o keyword en caso de que esta presentase ambigüedad.

Permite el enlace hacia LOD Cloud (Linkend Open Data Cloud), mediante el enlace de las entidades y las palabras claves encontradas en el texto al Dataset de la DBpedia y de este punto a cualquier repositorio enlazado que contenga estos recursos o enlaces.



Figura 12. Diagrama de secuencias módulo de Desambiguación y Enlace

Fuente: (Propio)

Descripción delos elementos del módulo.

**Desambiguación Main:** es subproceso principal para gestionar exclusivamente la comunicación entre los elementos para este módulo. Recibe los parámetro de los módulos anteriores: etiquetado y extracion.

**Buscar URI Recurso DBpedia:** mediante consultas Sparql hacía en endpoint de la DBpedia se consulta las entidades a las que correspondan las entidad y palabras claves extraídas con los módulos anteriores, teniendo en cuenta que muchas de estas tendrá más de un recurso de coincida en cuanto al “nombre” en DBpedia, teniendo así ambigüedad en cuanto a la correspondencia de la palabra (entidad o palabra clave) con un recurso de DBpedia.

**Extraer Abstract Recursos de DBpedia:** una vez que hemos obtenido los URIs de los recursos DBpedia, procedemos a consultar los *Abstracts* de estos recursos de la DBpedia que serán los parámetros de entrada para siguiente Algoritmo de Lesk en caso de una entidad o palabra clave (o palabras clave compuestas) se haya encontrado más de un recurso cuyo nombre coincide. Es impórtate contar con los *abstarct* de los elemento del texto que no presenten ambigüedad puesto que es “significado” ayudara a desambiguar los elementos ambiguos.

**Algoritmo Lesk:** en caso de una entidad, palabra clave o palabra claves tengan más de un recurso cuyo “nombre” coincida con DBpedia se aplica este algoritmo para determinar en base a los *abstracts* de los recursos la correspondencia más aproximada en base a las coincidencia de las palabra que contengan en los *abstracts*, los cuales han sido procesados para eliminar palabras que pueden alterar el resultado y cuyo significado no es relevante como artículos y verbos.

* 1. Servicios Web
  2. Implementación

CAPITULO 4

1. 1. Validación
      1. Comparación con servicios similares
   2. Pruebas

CAPITULO 5

1. 1. Conclusiones y Recomendaciones

# Bibliografía

Albahari, J., & Albahari, B. (2012). *C# 5.0 IN A NUTSHELL.* O’Reilly.

Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G., & Machina., L. (25 de 02 de 2014). *RDF 1.1 Turtle.* Obtenido de W3C Recommendation: http://www.w3.org/TR/2014/REC-turtle-20140225/

Berners-Lee, T. (01 de 2005). *Uniform Resource Identifier (URI): Generic Syntax.* Recuperado el 24 de 06 de 2014, de http://tools.ietf.org/html/rfc3986

Berners-Lee, T. (23 de Julio de 2006). Linked Data - Design Issues.

Bizer, C. (09 de 11 de 2009). *Dbpedia*. Recuperado el 10 de 06 de 2014, de The DBpedia Data Provision Architecture: http://wiki.dbpedia.org/Architecture

Bizer, C., & Cyganiak, R. (25 de 02 de 2014). *RDF 1.1 TriG.* Obtenido de W3C Recommendation: http://www.w3.org/TR/2014/REC-trig-20140225/

Carothers, G., & Seaborne, A. (25 de 02 de 2014). *RDF 1.1 N-Triples.* Recuperado el 25 de 06 de 2014, de W3C Recomendations: http://www.w3.org/TR/2014/REC-n-triples-20140225/

Clark, K. G., Feigenbaum, L., & Torres, E. (01 de 15 de 2008). *SPARQL Protocol for RDF.* Recuperado el 24 de 06 de 2014, de W3C Recommendation 15 January 2008: http://www.w3.org/TR/rdf-sparql-protocol/

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (Marzo de 1992). A Practical Part of Speech Tagger.

Cyganiak, R., Wood, D., & Lanthaler, M. (25 de 02 de 2014). *RDF 1.1 Concepts and Abstract Syntax.* Recuperado el 25 de 06 de 2014, de W3C Recommendation: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Lapuente, M. J. (08 de 12 de 2013). *HIPERTEXTO: EL NUEVO CONCEPTO DE DOCUMENTO EN LA CULTURA DE LA .* Recuperado el 24 de 06 de 2014, de Tesis doctoral. Universidad Complutense de Madrid.: http://www.hipertexto.info/

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., y otros. (2012). *DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia*. Obtenido de http://semantic-web-journal.net/system/files/swj499.pdf

McBride, B. (10 de 02 de 2004). *W3C Recommendation.* Recuperado el 26 de 06 de 2014, de RDF Primer: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

Miller, E. (Mayo de 1998). *An Introduction to the Resource Description Framework*. Recuperado el 27 de 06 de 2014, de http://www.dlib.org/dlib/may98/miller/05miller.html

Miller, E. (1998). Wiley Online Library. *Bulletin of the American Society for Information Science and Technology*, 15-19.

Pautasso, C., Zimmermann, O., & Leymann, F. (Abril de 2008). RESTful Web Services vs. “Big” Web Services: Making the Right Architectural Decision.

Prud'hommeaux, E., & Seaborne, A. (15 de 01 de 2008). *SPARQL Lenguaje de consulta para RDF.* Recuperado el 25 de 06 de 2014, de Recomendación del W3C de 15 de enero de 2008 : http://skos.um.es/TR/rdf-sparql-query/

Richardson, L., & Ruby, S. (2007). RESTful Web Services. En L. Richardson, & S. Ruby, *RESTful Web Services* (pág. 299). O'Reilly.

Ruby, L. R. (2007). *RESTful Web Services.*

Sandeep Chatterjee, j. W. (2004). *Developing Enterprise Web Services: An Architect's Guide.* Person Education Inc.

Schreiber, G., & Raimond, Y. (24 de 06 de 2014). *RDF 1.1 Primer.* Obtenido de W3C Working Group: http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. (2007). *In The semantic web.*

Tjong Kim Sang, E. F., & Buchholz, S. (Septiembre de 2000). Introduction to the CoNLL-2000 shared task: Chunking.

W3C. (01 de 2005). *Uniform Resource Identifier (URI): Generic Syntax.* Recuperado el 22 de 02 de 2014, de January 2005

W3C. (2013). *W3C.* Obtenido de http://www.w3.org/standards/semanticweb/ontology

Wood, D. (25 de 02 de 2014). *What’s New in RDF 1.1.* Obtenido de W3C: http://www.w3.org/TR/2014/NOTE-rdf11-new-20140225/

1. http://www.evolutionoftheweb.com [↑](#footnote-ref-1)
2. <http://www.w3.org/> [↑](#footnote-ref-2)
3. http://www.ietf.org/ [↑](#footnote-ref-3)
4. http://en.wikipedia.org/wiki/Request\_for\_Comments [↑](#footnote-ref-4)
5. http://www.ietf.org/rfc/rfc1738.txt [↑](#footnote-ref-5)
6. http://www.ietf.org/rfc/rfc3406.txt [↑](#footnote-ref-6)
7. [↑](#footnote-ref-7)
8. http://www.w3.org/TR/2014/REC-rdf-schema-20140225/ [↑](#footnote-ref-8)
9. http://www.foaf-project.org/ [↑](#footnote-ref-9)
10. http://dublincore.org/documents/dces/ [↑](#footnote-ref-10)
11. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#bib-RDF11-SCHEMA [↑](#footnote-ref-11)
12. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#bib-RDF11-SCHEMA [↑](#footnote-ref-12)
13. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/#dfn-rdf-compatible-xsd-types [↑](#footnote-ref-13)
14. http://www.w3.org/TR/sparql11-query/ [↑](#footnote-ref-14)
15. http://dbpedia.org/About [↑](#footnote-ref-15)
16. http://es.wikipedia.org/wiki/Crowdsourcing [↑](#footnote-ref-16)
17. <http://www.wikipedia.org/> [↑](#footnote-ref-17)
18. http://en.wikipedia.org/wiki/Help:Wiki\_markup [↑](#footnote-ref-18)
19. http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/ [↑](#footnote-ref-19)
20. http://es.wikipedia.org/wiki/Ambig%C3%BCedad [↑](#footnote-ref-20)
21. http://www.w3.org/ [↑](#footnote-ref-21)
22. Nombrado así en (Richardson & Ruby, 2007) [↑](#footnote-ref-22)
23. http://www.w3.org/TR/ws-gloss/#endpoint [↑](#footnote-ref-23)
24. http://www.thefreedictionary.com/self-contained [↑](#footnote-ref-24)
25. http://json.org/json-es.html [↑](#footnote-ref-25)