La web semántica

PID_00272096

Blas Torregrosa García

Tiempo mínimo de dedicación recomendado: 2 horas

Universitat Oberta de Catalunya

Blas Torregrosa García

Ingeniero en Informática y máster universitario en Seguridad de las Tecnologías de la Información y de las Comunicaciones (MISTIC) por la Universitat Oberta de Catalunya (UOC). Especializado en ciberseguridad. Profesor colaborador en el máster de Ciencia de Datos de la UOC y profesor asociado en la Universidad de Valladolid (UVA).

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Ferran Prados Carrasco (2020)

Primera edición: febrero 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

En sus años de vida, la World Wide Web (WWW), más conocida como **la web**, ya se ha convertido en una herramienta indispensable para la vida cotidiana de las personas, y ha llegado a ser el principal medio de comunicación mundial de información. Desde sus inicios, alrededor del año 1990 y fruto de la existencia de la red Internet, la web ha experimentado un crecimiento casi exponencial en términos de contenido.

Este contenido no es solo utilizado por personas, sino que cada vez hay una mayor necesidad de que pueda ser consumido por mecanismos automáticos. En este contexto y por este motivo los avances tecnológicos dirigidos al tratamiento automático del contenido son cada vez más necesarios y habituales. Para realizar tratamientos automáticos de datos y de información hay que identificar el significado de los datos de forma explícita, con el objetivo de ampliar la web con información semántica y convertirla en una web de datos.

1. Evolución de la web

En el año 1989 la World Wide Web (WWW, o simplemente, la web) surge a partir de una propuesta de Tim Berners-Lee para utilizar el hipertexto como mecanismo para intercambiar información. Actualmente la web es uno de los servicios más conocidos y populares de Internet.

La **web** es una red de páginas escritas en hipertexto y conectadas entre sí por medio de enlaces. Estas páginas están alojadas en diferentes servidores conectados entre ellos y que utilizan un protocolo (*HyperText Transfer Protocol* o HTTP) que permite descargar y consultar las páginas de hipertexto y los recursos que enlazan: imágenes, vídeo, audio, documentos, etc.

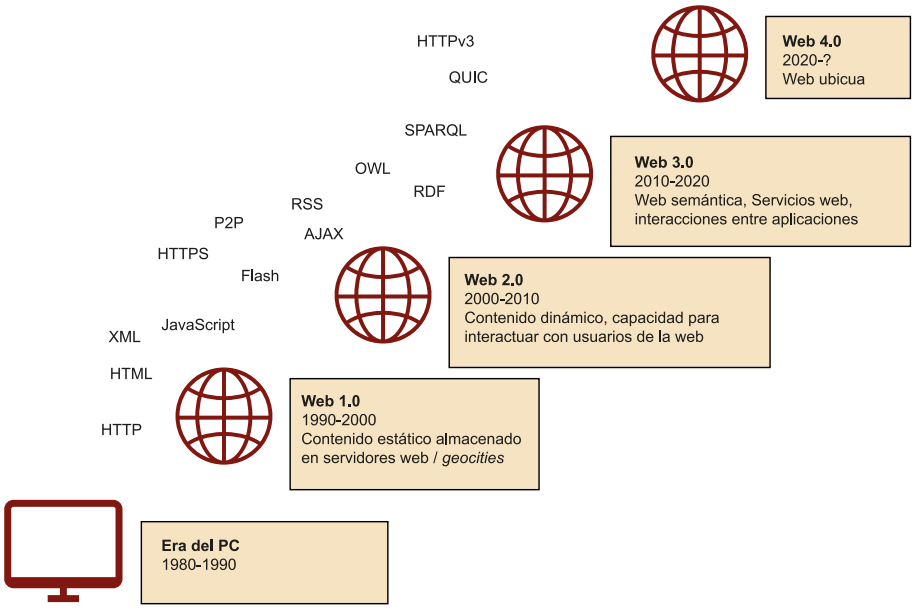
Cuando se quiere consultar una página web es necesario utilizar un software especial para esta función denominado **cliente web** o **navegador**,[1] que permite visualizar las páginas de hipertexto y guiar la navegación a través de los diferentes enlaces. En este sentido, el **hipertexto** incluye diferentes características que definen cómo se tiene que presentar la información de una página.

(1) En inglés, *browser*.

El **contenido** de las páginas web convencionales está diseñado para ser legible por personas, por lo que no es adecuado para un procesamiento automático y es muy poco eficiente cuando se busca información relacionada. Los conjuntos de datos en la web son silos de datos aislados que no están vinculados entre sí. Esta limitación puede abordarse organizando y publicando datos, utilizando formatos que agreguen estructura y doten de significado al contenido de las páginas web, vinculando los datos que estén relacionados entre sí. Los sistemas informáticos pueden «comprender» mejor este tipo de datos, lo que permite automatizar las tareas. Precisamente este es uno de los retos de la web semántica.

Para entender la evolución del contenido de la web, veremos cuál ha sido la evolución de su uso y de la tecnología subyacente.

Figura 1. Evolución de la web



1.1. La web 1.0

La web nació a principios de los años noventa como la unión de dos tecnologías ya existentes: el **hipertexto** e **Internet**.

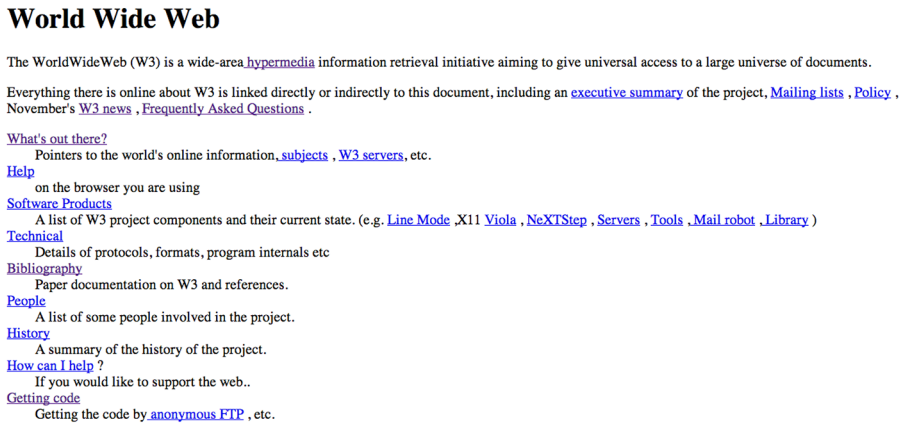
> **Web 1.0** es el término con el que se suelen denominar las primeras páginas web, caracterizadas por ofrecer información estática de manera unidireccional.

Es decir, el usuario que accede al contenido únicamente puede leerlo de manera pasiva, sin la posibilidad de contribuir en ningún caso a su ampliación o corrección. Por lo tanto el usuario es un simple consumidor de una web enfocada a la lectura.

Tecnológicamente aparece el lenguaje de marcas HTML (*HyperText Mark-Up Language*), lenguaje empleado para la creación de páginas web especialmente idóneo para enlazar contenidos web por medio de los hipervínculos (*hyperlinks*). También aparece un nuevo protocolo, el HTTP (*Hypertext Transfer Protocol*), y el sistema de localización de recursos URL (*Uniform Resource Locator*).

La popularidad alcanzada por la web a partir de 1994 rebasó las fronteras del intercambio científico y permitió el inicio de las primeras aplicaciones comerciales. Muchas de las compañías del comercio electrónico, como Amazon, eBay o Yahoo!, se fundaron entre 1994 y 1995.

Figura 2. Primera página web



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
Pointers to the world's online information, subjects , W3 servers, etc.
Help
on the browser you are using
Software Products
A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
Details of protocols, formats, program internals etc
Bibliography
Paper documentation on W3 and references.
People
A list of some people involved in the project.
History
A summary of the history of the project.
How can I help ?
If you would like to support the web..
Getting code
Getting the code by anonymous FTP , etc.

Fuente: home.cern/science/computing/birth-web

Figura 3. Primeras versiones de Google, creada en 1996

En el año 1995 Microsoft lanza el sistema operativo Windows 95 que incluye el navegador Internet Explorer. La competencia entre Internet Explorer y Netscape (navegador por entonces más popular) da lugar a la primera guerra de navegadores, en la que los navegadores compiten entre sí incluyendo cada vez más características y forzando los límites del lenguaje HTML.

1.2. La web 2.0

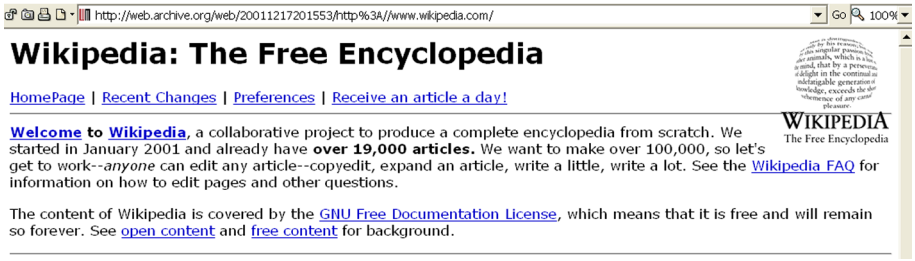
Casi una década más tarde aparece el concepto de **web 2.0**, también conocida como **web colaborativa**. Se considera una evolución de la web 1.0, con el objetivo de permitir una mayor interacción con el contenido, es decir, dotar a las páginas de mecanismos para la colaboración de los usuarios en la transformación y creación de nuevo contenido. De esta manera, el usuario pasa a ser también productor, y se consigue una comunicación bidireccional con la web.

Para hacer esto posible, es necesario que tecnológicamente se permita la modificación de contenidos con el mínimo conocimiento técnico. Por ello aparecen herramientas web como los blogs o las wikis, entre otros. Esto representa la convergencia entre medio de comunicación y contenido, y en este contexto aparecen las primeras redes sociales, como MySpace, YouTube, etc.

Figura 4. Diseño original y nombre de Thefacebook en 2004

Fuente: Wikipedia.

Figura 5. Diseño de Wikipedia en 2001



Fuente: Wikipedia.

> La **web 2.0** tiene dos características esenciales: la consolidación de las redes sociales y el acercamiento al tiempo real.

La web social se popularizó gracias al éxito de Facebook y de otras redes sociales. Y la web en tiempo real podría estar representada por Twitter, que comenzó su andadura en el año 2006 como un sistema de mensajes cortos, parecidos a los SMS (menos de 140 caracteres), para la web.

Por otro lado, en paralelo se avanza en tecnologías para el cliente: se consolida JavaScript en los navegadores y las técnicas de acceso asíncrono como AJAX (con *frameworks* como jQuery), y también el estándar de intercambio de datos JSON.

1.3. La web 3.0

La **web 3.0** se conoce generalmente como la **web semántica** o también **la web de lectura-escritura-ejecución**.

La web 3.0 descentraliza servicios como la búsqueda, las redes sociales y las aplicaciones de mensajería instantánea que dependen de una sola organización para funcionar. La web semántica y los servicios web son los principales componentes de la web 3.0.

El objetivo de la web semántica es transformar la actual web sintáctica (una web de documentos), en la que la unidad de información es el documento, en una **web de datos**, en la que la unidad de información sea el dato. Para conseguirlo, primero hay que dotar de significado a los recursos web, es decir, se tiene que clasificar, estructurar y anotar semánticamente cada recurso para que pueda ser interpretado por las aplicaciones que lo tendrán que procesar.

Los servicios web son un método de comunicación entre sistemas informáticos a través de una red de comunicaciones, y tal comunicación se realiza de manera estandarizada (mediante XML, JSON, SOAP, etc.) que permita la integración de aplicaciones web heterogéneas.

Desde el año 2011 se detecta un uso masivo de dispositivos móviles para acceder a la web.

1.4. La web 4.0

La **web 4.0** es el próximo gran avance que se centrará en ofrecer un comportamiento más inteligente y se basará en explotar toda la información que ahora mismo contiene, pero de una forma más natural y efectiva.

Actualmente, los buscadores siguen siendo elementos esenciales. Lo que propone la web 4.0 es mejorar esta experiencia mediante el uso de tecnologías que permitirían un nivel de interacción más completo y personalizado, usando toda la información que damos y existe en la web. Todo ello se fundamentará en cuatro pilares:

1) La comprensión del lenguaje natural y tecnologías de conversión de voz en texto y viceversa.

2) Sistemas de comunicación de máquina a máquina (M2M).

3) Uso de la información de contexto como el posicionamiento GPS o el ritmo cardiaco detectado por un *wearable*, dispositivo móvil, etc.

4) Modelo mejorado de interacción con el usuario.

La web 4.0 sería la unión de la web semántica, la inteligencia artificial y la voz como forma de comunicación. El objetivo es una web ubicua cuyo propósito primordial será el de unir las inteligencias que se comunican entre sí para generar la toma de decisiones.

2. La web semántica

2.1. ¿Qué es la web semántica?

> «La web semántica es una extensión de la actual web en la que la información tiene un significado bien definido y permite que personas y ordenadores cooperen».
>
> Tim Berners-Lee, James Hendler, Ora Lassila (2001). *The Semantic Web, Scientific American* (vol. 5, núm. 284, págs. 34-43).

Una definición de web semántica comienza con la definición de la **semántica**, es decir, el significado. Las páginas web están llenas de datos y etiquetas asociadas. La mayoría de las etiquetas representan instrucciones de formato, como `<H1>` para indicar un encabezado. Semánticamente sabemos que las palabras rodeadas con `<H1>` son más importantes que otro texto que no lo esté.

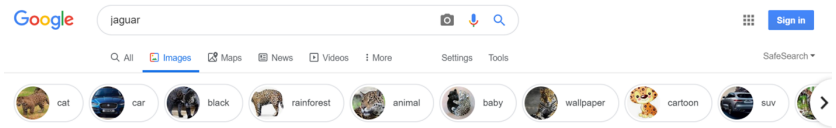
Semántica

La palabra *semántica* también se usa en la web en más contextos. Por ejemplo, en HTML5 hay elementos estructurales semánticos (como `section` que permite agrupar por temas), aunque esta expresión se refiere al «significado» de los elementos. Es decir, no hay que confundir la semántica de los elementos del marcado con la semántica (la capacidad de procesamiento por las máquinas) de las anotaciones utilizadas en la web semántica.

Las páginas web se basan en lenguajes de marcado (HTML) para determinar la estructura del documento; las hojas de estilo (CSS), para la apariencia, y *scripts* (JavaScript) para el comportamiento, aunque el contenido sigue estando orientado a ser comprensible básicamente por personas.

Si se busca «Jaguar» en la web, por ejemplo, los algoritmos de los motores de búsqueda no saben si nos referimos a una marca de automóviles de lujo o a un felino depredador sudamericano.

Figura 6. Búsqueda



Fuente: Google.

Por defecto, los encabezados, textos, enlaces y otros elementos de las páginas web no tienen sentido para los ordenadores. Los navegadores simplemente muestran los documentos web, aunque solo el cerebro humano puede interpretar el significado.

El concepto de datos entendibles por máquinas no es nuevo y no se limita a la web. Por ejemplo, las tarjetas de crédito o los códigos de barras contienen datos legibles por humanos y por máquinas.

Incluso los documentos XML, que tienen unas reglas sintácticas rigurosas, tienen sus limitaciones cuando son procesados por máquinas. Por ejemplo, si se define una entidad XML con el marcado `<CMD>` y `</CMD>`, ¿qué significa realmente CMD? Puede referirse a un comando del sistema o a un centro de mando distribuido o a un camarote doble.

Para hacer que los contenidos sean procesables sin ambigüedades por aplicaciones informáticas es necesario añadir a las páginas web datos organizados (estructurados), como anotaciones o como metadatos, que hagan que estén vinculados a otros datos estructurados relacionados.

La ventaja de añadir anotaciones semánticas a las páginas web es que las personas podrán seguir navegando por los documentos web, mientras que las aplicaciones informáticas podrán procesar esas anotaciones para clasificar las entidades de datos, descubrir relaciones lógicas entre entidades, crear índices, etc.

La **web semántica** es un conjunto de estándares y de buenas prácticas que permite compartir datos en la web y su semántica para su uso por aplicaciones informáticas.

Web semántica

La web semántica está impulsada por el Consorcio World Wide Web (W3C).

Tabla 1. La web tradicional frente a la web semántica

Características	Web tradicional	Web semántica
Componente fundamental	Contenido no estructurado, lenguaje natural	Contenido estructurado
Audiencia	Personas	Máquinas/Aplicaciones
Enlaces	Indican localización	Indican localización y significado
Basado en	Sintaxis	Semántica

2.2. Comprendiendo el contenido de la web

Cuando se accede a un contenido en la web, no importa en qué idioma esté, una persona o un sistema informático tiene que considerar las siguientes cuestiones:

- ¿Qué información es importante y cómo puedo saberlo?
- ¿Qué es información y qué es publicidad?
- ¿Qué información está relacionada con el contenido?
- ¿Qué significa la información?

Las personas tienen conocimiento del contexto y experiencia para resolver estas cuestiones. Para los sistemas informáticos es una tarea compleja filtrar y obtener información de la web para determinar lo que es importante. Es decir, para entender el significado de la información de la web.

2.2.1. La importancia del significado

Significado, comprensión y entendimiento son tres conceptos interrelacionados. El **entendimiento** es la capacidad de comprender el significado de la información, mientras que la **comprensión** es hacer propio lo que se entiende y asimilarlo. La información se envía mediante un mensaje desde un remitente hasta un receptor usando un lenguaje concreto.

El receptor del mensaje entiende la información si el receptor interpreta correctamente la información. La **correcta interpretación de la información** depende de:

- **Sintaxis**.[2] En Gramática, es el estudio de los principios y procesos por los que las oraciones están bien construidas en un determinado idioma. En lenguajes formales, la sintaxis es un conjunto de reglas por el que se generan expresiones correctas (bien formadas) con un conjunto de símbolos (alfabeto). En informática, la sintaxis regula la estructura de los datos, es decir, las reglas de lo que está permitido y lo que no.

(2)Del griego, *sin* (junto) y *taxis* (orden) que significa 'coordinación'.

- **Semántica**.[3] Es la parte de la lingüística centrada en el sentido y significado del lenguaje y es el estudio de la interpretación de los símbolos usados por una comunidad en unas circunstancias particulares o contexto. La semántica también utiliza reglas de sintaxis para determinar el sentido y significado de conceptos complejos que derivan de conceptos simples. La semántica de un mensaje depende del contexto y de la pragmática.

(3)Del griego, *semantikos* ('explicación de lo que significa').

- **Contexto**.[4] En una expresión, denota a todo lo adyacente a un símbolo (concepto) respecto a las relaciones con expresiones (conceptos) adyacentes y otros elementos relacionados. El contexto se refiere a todos los elementos de cualquier tipo de comunicación que determinan la interpretación del contenido comunicado. Podemos distinguir entre contextos **generales** (lugar, tiempo, etc.) y contextos **personales o sociales** (relación entre el remitente y el receptor).

(4)De latín, *contextus* ('conexión', 'entretejido').

- **Pragmática**.[5] Refleja la intención del uso del lenguaje para comunicar un mensaje, es decir, la finalidad del remitente. En lingüística, es el estudio del uso del idioma en diferentes situaciones. La pragmática estudia las formas en las que el contexto contribuye al significado.

(5)Del griego, *pragmatikos* ('acción').

- **Experiencia**.[6] Considera toda la información aprendida y puesta en un contexto, es decir, conocimiento mundial o sentido común que influye en cómo se interpreta la información.

(6)Del latín, *experientia* ('acontecimiento vivido').

En resumen, para una comunicación eficaz es necesario que la información sea correctamente transmitida (sintaxis) y el significado (semántica) de la información transmitida sea interpretado correctamente. El entendimiento del mensaje dependerá del contexto de remitente y receptor así como de la intención del remitente. Y el contexto dependerá de la experiencia que tengan remitente y receptor.

2.2.2. De la web tradicional a la web de datos

El problema de la web tradicional es que no tiene una semántica explícita, puesto que la mayoría de la información que se transfiere en la web está codificada en lenguaje natural o dentro de contenidos multimedia, como imágenes, vídeos, audios, etc. Por lo tanto el significado está oculto y esto dificulta que los sistemas informáticos puedan entender el contenido de la web.

La web de datos o web semántica se considera una actualización de la web de documentos tradicional.

La web de datos considera la web como una enorme base de datos descentralizada (base de conocimiento) con datos accesibles para los sistemas informáticos.

Para conseguir una web de datos hace falta una condición previa: el contenido de la web tiene que ser leído e interpretado correctamente (entendido) por máquinas. Existen dos enfoques para obtener esto:

1) Motores de búsqueda con procesamiento del lenguaje natural[7] que utilizan las tecnologías de recuperación de información que tratan de entender el contenido de la web mediante estadísticas y aprendizaje automático. El objetivo es extraer la semántica implícita en la web.

(7) *Natural Language Processing* (NLP).

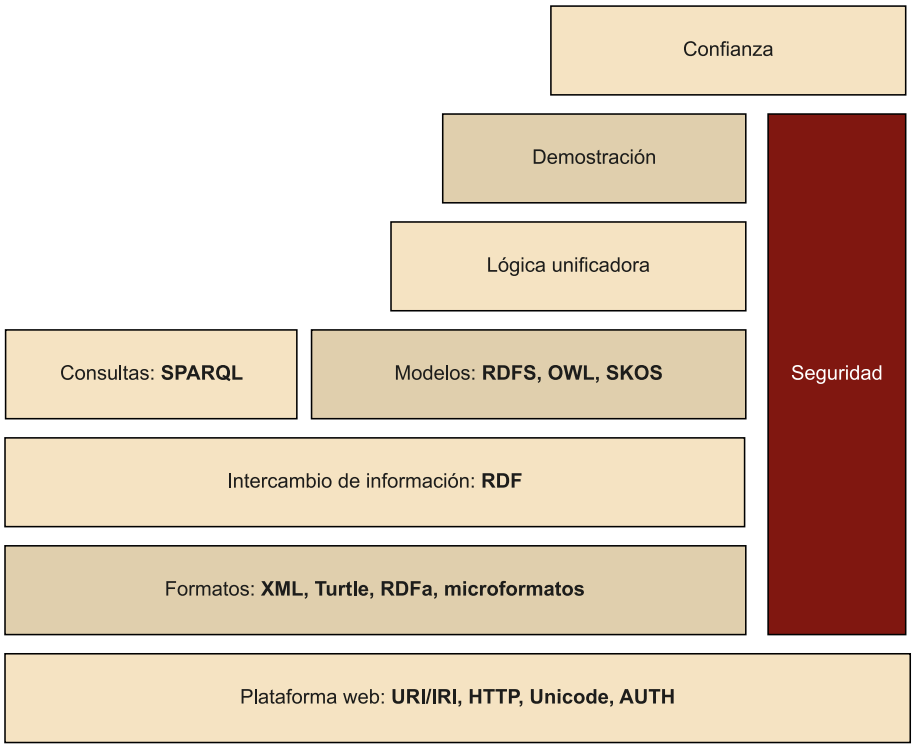
2) Tecnologías de la web semántica en las que el contenido en lenguajes naturales se anota explícitamente con metadatos semánticos. Los metadatos semánticos codifican el significado del contenido y puede entonces ser leído e interpretado correctamente por máquinas.

2.3. Arquitectura de la web semántica

En el año 2000 Tim Berners-Lee propuso un modelo de capas para esquematizar el desarrollo futuro de la web semántica. Su modelo se bautizó como **la tarta de la web semántica**.[8] Más adelante, en 2006, se publicó una nueva versión de la tarta.

(8) En inglés, *Semantic Web Layer Cake*.

Figura 7. La tarta de la web semántica



En este modelo se distinguen las siguientes capas:

1) Plataforma web: la primera capa se refiere a las tecnologías de infraestructura más básicas, como el HTTP (*HyperText Transfer Protocol*), Unicode para codificación de caracteres internacionales, URI para identificar recursos y mecanismos de autenticación para seguridad.

2) Formatos: la siguiente capa se refiere a los formatos para representar o serializar la información de las capas superiores. Se puede representar todo en XML, Turtle, RDFa o microformatos.

3) Intercambio de información: es el bloque principal que desarrolla la infraestructura que permitirá la descripción de recursos. En esta capa el lenguaje de referencia es RDF (*Resource Description Framework*).

4) Modelos: en esta capa de organización de la información se agrega semántica a los datos definidos con RDF mediante vocabularios definidos en RDFS (RDF *Schema*) o SKOS (*Simple Knowledge Organization System*), o mediante la definición de ontologías para los distintos dominios con OWL (*Web Ontology Language*), que es el lenguaje recomendado para la definición de ontologías. Estos modelos son representaciones del conocimiento.

5) Consultas: se propone un lenguaje de consultas para la web de datos que desempeñe un papel similar al del lenguaje SQL para las bases de datos relacionales. Ese lenguaje se denomina **SPARQL** (*SPARQL Protocol and RDF Query Language*).

6) Lógica/Demostración: en estas dos capas se identificaban las cuestiones de inferencia que habría que desarrollar para generar nuevo conocimiento a partir de la web de datos.

7) Confianza: si se pretende que se puedan realizar tareas de forma autónoma a partir de la información publicada en la web, es necesario desarrollar una infraestructura que permita asegurar su fiabilidad. Dicha infraestructura se apoyaría en firmas digitales que permitirían identificar la autoría de las publicaciones.

Este modelo en capas ha recibido críticas. Lo más controvertido de la tarta de la web semántica es que se pretende que cada capa de un nivel se base en el nivel anterior. Por un lado, existen algunas dificultades técnicas para asegurar que cada capa superior se base en la capa inferior. Y por otro, algunas tecnologías de la web semántica todavía siguen en desarrollo.

Por todo ello, la tarta no debe entenderse de forma literal, sino como un mapa conceptual de las tecnologías implicadas.

2.3.1. Nombrando las cosas: URL, URI y URN

Los **recursos web** se pueden localizar mediante direcciones IP que son únicas. Sin embargo las direcciones IP son bastante difíciles de recordar. Por ello se utilizan los **nombres de dominio**, que siguen unas reglas de sintaxis. Los nombres de dominio convencionales no pueden contener caracteres acentuados o no alfanuméricos. Con la introducción de los nombres de dominio internacionales (IDN)[9] es posible usar nombres en varios idiomas y con diferentes alfabetos.

[9]Acrónimo del inglés *Internationalized Domain Names*.

IP

IPv4 utiliza direcciones de 32 bits (4 números de 8 bits separados por punto) que limitan el espacio de direcciones a 4.294.967.296 (2^{32}) direcciones. Mientras que una dirección IPv6 está formada por 128 bits.

Los **URI** (*Uniform Resource Identifier*), cuyo subconjunto más conocido son los URL (*Uniform Resource Locators*), proporcionan el mecanismo para identificar de forma unívoca cualquier recurso de la web: documentos, imágenes, vídeos, etc. En la web semántica, los URI tendrán además la función de identificadores de objetos del mundo real. Cualquier persona u objeto podrá ser identificado mediante un URI.

Reglas de sintaxis

Reglas de sintaxis definidas en las RFC 1035, RFC 1123 y RFC 2181.

Figura 8. Formato de URI

Los URI tienen la siguiente estructura:

- **Esquema**: es un nombre que se refiere a la especificación para asignar identificadores como owl: o rdf:, y también puede identificar el protocolo de acceso al recurso, como http:, ftp:, etc.

- **Autoridad**: es un nombre jerárquico que representa el nombre de dominio en Internet. Comienza por el símbolo //.

- **Ruta**: es una secuencia de segmentos separados por / y en forma jerárquica, similar a los directorios en sistemas de fichero.

- **Consulta**: es una parte que comienza por ?, aporta información opcional, normalmente mediante pares atributo=valor separados por el símbolo &.

- **Fragmento**: es una parte también opcional que comienza por # y permite identificar una parte o recurso secundario dentro de un recurso principal.

En muchas ocasiones se utiliza los URI como sinónimo de URL, aunque URI es un término más amplio. Los URI se pueden clasificar como URL, como **URN** (*Uniform Resource Names*), o ambos. Un URN define la identidad de un recurso, mientras que el URL proporciona un método para encontrarlo (incluyendo el protocolo y la ruta).

Como la notación de los URI puede ser bastante larga, existen también los CURIE (*Compact* URI), que es una notación abreviada para los URI. La sintaxis consta de tres partes:

1) El prefijo (dbpedia como prefijo de http://dbpedia.org/resource/)
2) El carácter dos puntos (:)
3) La referencia

Planeta Venus

Por ejemplo, para referencia el planeta Venus está el URI http://dbpedia.org/resource/Venus que se puede abreviar como dbpedia:Venus.

Además existe el **IRI** (*Internationalized Resource Identifier*), que es una extensión del URI, que permite utilizar el juego de caracteres Unicode, con lo que se pueden incluir caracteres chinos, kanji japoneses, árabes o cirílicos para identificar recursos.

Muchas veces ocurre que los recursos cambian de localización o de dominio. Aunque las direcciones web se pueden redirigir (generalmente usando la redirección con HTTP 302) a la nueva dirección, esto puede causar problemas. Una opción es usar una localización persistente en la red; para ello habrá que utilizar **PURL** (*Persistent Uniform Resource Locator*).

2.3.2. Formatos

En un nivel superior se encuentran los documentos y su estructuración lógica. Originalmente, XML (*eXtensible Markup Language*) constituye la base sintáctica de la web semántica y sobre la que se apoyan el resto de las capas. XML es un metalenguaje que permite definir diferentes lenguajes de etiquetado validándolos mediante definiciones de documentos o DTD (*Document Type Definitions*) y también con XML *Schemas*. Sin embargo tiene el inconveniente de ser un lenguaje muy verboso y complicado de leer por las personas.

Otro formato habitual es Turtle (*Terse RDF Triple Language*), que fue desarrollado expresamente para ser legible por las personas. Se trata de un formato de texto donde aparecen codificadas las tripletas del lenguaje RDF.

Últimamente hay dos formatos que han sido de interés para el desarrollo de la web semántica: JSON-LD y microformatos. JSON-LD (*Javascript Object Notation for Linked Data*) es un formato más legible para las personas y que utiliza JSON de conjuntos de pares atributo-valor. Un microformato (abreviado como µF) es un método para agregar semántica usando HTML con un marcado específico.

2.3.3. El lenguaje RDF

RDF (**Resource Description Framework**) es un lenguaje de etiquetado, creado mediante XML, que define un modelo de datos para describir recursos (cualquier objeto identificable por un URI). Para ello utiliza enunciados en forma de tripletas sujeto-predicado-objeto (recurso-propiedad-valor), donde sujeto y predicado son URI y el objeto puede ser una URI o un valor literal. El predicado (propiedad) describe la relación entre el sujeto y el objeto. El lenguaje RDF es el equivalente a HTML (*HyperText Markup Language*) en la web convencional.

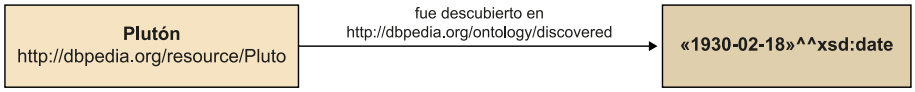
Si quisiéramos expresar la frase en lenguaje natural «Plutón fue descubierto en 1930» mediante tripletas RDF sería:

Tabla 2. Ejemplo de tripletas RDF

	Modelo de datos RDF	Tripleta RDF	Tipo
Sujeto	Plutón	http://dbpedia.org/resource/Pluto	URI
Predicado	Descubierto	http://dbpedia.org/ontology/discovered	URI
Objeto	1930	«1930»	Literal

RDF es también un **grafo** dirigido, en el que los nodos son el sujeto y el objeto, y los arcos son los predicados que los conectan.

Figura 9. Ejemplo de grafo RDF



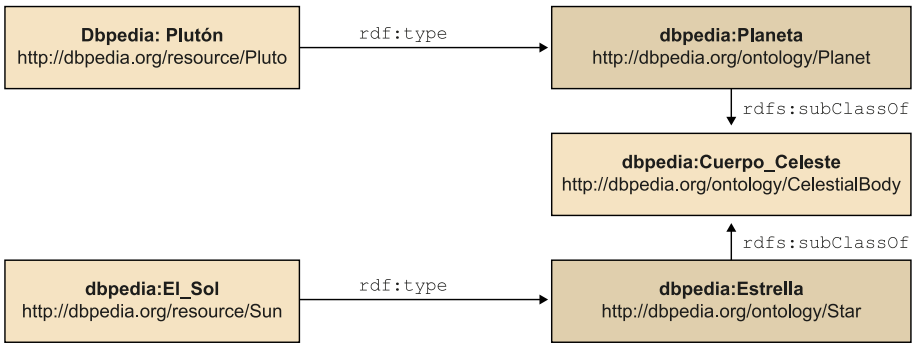
2.3.4. RDF Schema

RDF **Schema** (RDFS) es un vocabulario RDF que nos permite describir recursos mediante una orientación a objetos similar a la de muchos lenguajes de programación como Java. Para ello proporciona un mecanismo para definir clases, objetos y propiedades, así como relaciones entre clases y propiedades. También restricciones de dominio y rango sobre las propiedades.

RDFS permite la definición de clases con `rdfs:Class` y la instanciación de clases en RDF con `rdf:type`. Existen más definiciones como `rdf:Property`, `rdfs:domain`, `rdfs:range`, `rdfs:Literal` o `rdfs:Resource`.[10] RDFS también define relaciones jerárquicas con `rdfs:subClassOf` y `rdfs:subPropertyOf`.

(10)Todo en el modelo RDF son recursos.

Figura 10. Ejemplo de esquema con RDFS



2.3.5. Ontologías

Las ontologías se utilizan como modelo de **representación de conocimiento** en el campo de la inteligencia artificial.

Se puede definir **ontología** como una especificación explícita y formal de un concepto dentro de un determinado dominio de interés.

En nuestro caso, una ontología se utiliza como un artefacto que define:

- Un **vocabulario compartido** que describe un determinado dominio.
- Un **conjunto de hipótesis** sobre los términos de dicho vocabulario; generalmente se utiliza un **lenguaje formal** manipulable automáticamente.

Las ontologías se pueden usar para buscar, consultar, indexar y administrar metadatos, y para mejorar la interoperabilidad de aplicaciones y bases de datos.

OWL (*Web Ontology Language*) es un lenguaje que extiende RDF y RDFS, y que permite la construcción de clases complejas a partir de otras definiciones de clases, así como el encadenamiento de propiedades.

3. DBpedia

La DBpedia es un proyecto para extraer de la Wikipedia datos estructurados.

La versión en inglés de la base de conocimiento de la DBpedia de junio de 2018 describe 6,6 millones de entidades, de las cuales 4,9 millones tienen resúmenes, 1,9 millones tienen coordenadas geográficas y 1,7 millones de representaciones.

En total, 5,5 millones de recursos se clasifican en una ontología consistente, que consta de 1,5 millones de personas, 840.000 lugares (incluyendo 513.000 lugares habitados), 496.000 obras (incluyendo 139.000 álbumes de música, 111.000 películas y 21.000 videojuegos), 286.000 organizaciones (incluidas 70.000 compañías y 55.000 instituciones educativas), 306.000 especies, 58.000 plantas y 6.000 enfermedades.

3.1. De la Wikipedia a la DBpedia

La **Wikipedia** se ha convertido en una fuente de información de referencia sobre conceptos, hechos, ciencia y cultura utilizada por millones de usuarios. El proyecto **DBpedia** surge sobre la base de la formalización del conocimiento de los artículos de Wikipedia.

Los artículos de Wikipedia no solo incluyen contenido textual; buena parte de ellos también contiene una gran cantidad de información estructurada mediante fichas descriptivas (*infoboxes*).

Una ***infobox*** es una plantilla wiki en la que se define una estructura de datos común y su representación visual para determinados tipos de artículos (personas, ciudades, películas, etc.).

Figura 11. Página de Wikipedia con *infobox* resaltado

Plutón (planeta enano)

Plutón, designado **(134340) Pluto**, es un planeta enano del sistema solar situado a continuación de la órbita de Neptuno. Su nombre se debe al dios mitológico romano Plutón (Hades según la mitología griega). En la Asamblea General de la Unión Astronómica Internacional celebrada en Praga el 24 de agosto de 2006 se creó una nueva categoría llamada plutoide, en la que se incluye a Plutón. Es también el prototipo de una categoría de objetos transneptunianos denominada plutinos. Plutón posee una órbita excéntrica y altamente inclinada con respecto a la eclíptica, que recorre acercándose en su perihelio hasta el interior de la órbita de Neptuno. Asimismo posee también cinco satélites: Caronte, Nix, Hidra, Cerbero y Estigia,[3] [4] los cuales son cuerpos celestes que comparten esa misma categoría.

Su gran distancia al Sol y a la Tierra, unida a su reducido tamaño, impide que brille por encima de la magnitud 13,8 en sus mejores momentos (perihelio orbital y oposición), por lo cual solo puede ser apreciado con telescopios a partir de los 200 mm de abertura, fotográficamente o con cámara CCD. Incluso en sus mejores momentos aparece como astro puntual de aspecto estelar, amarillento, sin rasgos distintivos (diámetro aparente inferior a 0,1 segundos de arco). No fue hasta el año 2015 cuando la sonda espacial New Horizons pasó sobre el planeta y permitió apreciar por primera vez de forma nítida su aspecto real.

Plutón fue descubierto el 18 de febrero de 1930 por el astrónomo estadounidense Clyde William Tombaugh (1906-1997) desde el Observatorio Lowell en Flagstaff, Arizona, y fue considerado el noveno y más pequeño planeta del sistema solar por la Unión Astronómica Internacional y por la opinión pública desde entonces hasta 2006, aunque su pertenencia al grupo de planetas del sistema solar fue siempre objeto de controversia entre los astrónomos. Durante muchos años existió la creencia de que Plutón era un satélite de Neptuno que había dejado de ser satélite por el hecho de alcanzar una segunda velocidad cósmica. Sin embargo, esta teoría fue rechazada en la década de 1970.[5]

Tras un intenso debate, y con la propuesta de los astrónomos uruguayos Julio Ángel Fernández y Gonzalo Tancredi ante la Asamblea General de la Unión Astronómica Internacional en Praga, República Checa, se decidió por

Fuente: Wikipedia.

De esto precisamente surge DBpedia, intentando convertir el conocimiento de las *infoboxes* de la Wikipedia en conocimiento formalizado mediante una ontología que aplica los principios y tecnologías de datos abiertos y enlazados.

3.2. Wikidata

Wikidata es un concentrador de datos estructurados en el que cada entidad se representa mediante un IRI, en el que también se recogen los enlaces a los artículos equivalentes de Wikipedia en diferentes idiomas.

Wikidata

Así, la entidad Q339 referencia al planeta Plutón del que existen artículos en inglés, español, catalán y gallego.

Aunque pudiera parecer que DBpedia y Wikidata son proyectos muy parecidos, que producen datos estructurados derivados de los artículos de Wikipedia, existen diferencias, desde la identificación de los recursos (URI/IRI para la DBpedia, mientras que Wikidata usa identificadores numéricos independientes del idioma) hasta la estructura interna (RDF para la DBpedia, mientras Wikidata desarrolla su propio modelo de datos).

Bibliografía

Allemang, D.; Hendler, J. (2011). *Semantic Web for the working ontologist* (2.ª ed.). Massachusetts: Morgan Kaufmann.

Saorín, T.; Pastor-Sánchez, J. A. (2018). «Wikidata y DBpedia: viaje al centro de la web de datos». *Anuario ThinkEPI* (n.º 12, págs. 207-214).

Sikos, L. F. (2015). *Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data*. Apress.

Taylor, J.; Evans, C.; Segaran, T. (julio 2009). *Programming the Semantic Web*. O'Reilly Media.