IES sujeta a verificación
del Ministerio de Educación

Caracterización de la Web Colombia mediante la herramienta WIRE

Arturo Eraso Torres

Grupo de Investigación Tecnofilia

Caracterización de la Web Colombia mediante la herramienta WIRE

Eraso Torres, Arturo
Caracterización de la web Colombia mediante la herramienta WIRE / Arturo Eraso Torres. -- 1 ed. -- San Juan de Pasto: Institución Universitaria Centro de Estudios Superiores María Goretti, 2016.
76 p.: il. ; 18 cm.

ISBN: 978-958-56064-0-1
e-ISBN: 978-958-56064-1-8
DOI: 10.15658/CESMAG16.010001

1. 2. WORLD WIDE WEB (SERVICIO DE INFORMACIÓN SOBRE REDES. II. Tít.

CDD 004.68
20. Ed.
CEP - Institución Universitaria Centro de Estudios Superiores María Goretti CESMAG. Biblioteca Remigio Fiore Fortezza.

Caracterización de la web Colombia mediante la herramienta WIRE

Primera edición, diciembre de 2016

Carrera 20A No.14-54
Tel: +572 – 7216535 Ext. 377 - 218
E-mail: editorial@iucesmag.edu.co
Website: www.iucesmag.edu.co/editorial
San Juan de Pasto, Nariño, Colombia
CP: 520003

Grupo de investigación TECNOFILIA
Facultad de Ingeniería
Programa de Ingeniería de Sistemas
Carrera 20A 14-54
Tel: +572 – 7216535 Ext. 240 - 232
E-mail: aeraso@iucesmag.edu.co

ISBN: 978-958-56064-0-1
e-ISBN: 978-958-56064-1-8
DOI: 10.15658/CESMAG16.010001

Rector:
Fray Hugo Ariel Osorio Osorio, OFM, Cap.

Directora editorial:
María Eugenia Córdoba

Edición:
Emma del Pilar Rojas Vergara
Diego Martínez Hernández
Pbro. Emilio Acosta Díaz

Edición impresa y digital
Impreso y hecho en Colombia
Printed and made in Colombia

Diseño de cubierta y diagramación:
Editorial Institución Universitaria CESMAG
D.G. Ana Cristina Benavides Erazo, acucristi14@gmail.com

Impresión: Compugráficas Pasto.

APA: Eraso Torres, A. (2016). *Caracterización de la web Colombia mediante la herramienta WIRE*. Pasto, Colombia: Editorial Institución Universitaria CESMAG,doi: 10.15658/CESMAG16.010001

CONTENIDO

CONTENIDO

CONTENIDO

LISTA DE FIGURAS

CONTENIDO

LISTA DE CUADROS

INTRODUCCIÓN

La tecnología evoluciona continuamente. El desarrollo de las redes de computadoras ha posibilitado una nueva forma de comunicación y ahora están permitiendo la consolidación de la llamada comunidad virtual. Hoy todo es posible a través de una tecnología denominada Web, un servicio inventado por Timothy Berners-Lee, quien lo desarrolló para un propósito específico de investigación y que luego daría como resultado una gama de servicios que llevarían al cambio de muchos paradigmas económicos, sociales, de comunicación y convivencia, que han conllevado a lo que hoy se conoce como Internet.

En este orden de ideas, el crecimiento de esta ola tecnológica no se ha hecho esperar y todos los actores de los procesos han querido hacer parte de ella: las universidades que por su vocación investigativa han sido las llamadas a abrir caminos que han permitido conocer más acerca del tema, luego las grandes empresas que han vislumbrado una oportunidad de negocio en la implementación de esta tecnología y, por último, las pequeñas empresas y emprendedores independientes que quieren una oportunidad de crecer y consolidarse.

Esta globalización que se extiende cada día más, ha permitido ver la Web como un todo, sin ningún tipo de frontera como el idioma o la división geográfica; sin embargo, es necesario entender que ésta se encuentra caracterizada por el desarrollo general de una región o país y abarca una gran variedad de aspectos como: las políticas de gobierno, las capacidades tecnológicas y los conocimientos específicos.

La caracterización de un espacio de la Web da a conocer valiosa información tecnológica que permite comprender su estructura, composición y demás particularidades.

Se han realizado diferentes estudios, algunos como en Brasil por Marco Modesto, et. al[1]., España por Ricardo Baeza-Yates, Carlos Castillo y Vicente López[2],

[1] MODESTO, Marco, et. al. Um novo retrato da web brasileira. In Proceedings of XXXII SEMISH, São Leopoldo, Brazil, 2005. p. 13.

[2] BAEZA-YATES, Ricardo; CASTILLO, Carlos y LOPEZ, Vicente. Características de la web de España. El Profesional de la Información, 2006. p. 16.

Argentina por Gabriel Tolosa, et. al[3]., Chile realizado por Carlos Castillo, Eduardo Graells y Ricardo Baeza-Yates[4]; de ahí que el objetivo del presente trabajo investigativo sea el de caracterizar el espacio Web de Colombia, partiendo desde el componente básico que tiene la Web: las páginas, hasta llegar a analizar los sitios y dominios desde diferentes puntos de vista, todo con el único propósito de entender su estructura, sus componentes, sus particularidades y similitudes con las Web de otros países.

El siguiente es un resumen de los contenidos del presente documento:

El Capítulo 1. La Web: Corresponde a los referentes y conceptos que se tuvieron en cuenta para proceder al desarrollo investigativo, es de acotar que es fundamental para el lector que quiere comprender la idea investigativa, las herramientas que se utilizaron, las definiciones que dan pie al uso de términos y los referentes sobre los cuales se soporta el estudio, con el fin de presentar unos datos relacionados con la Web de Colombia de forma consistente.

En el Capítulo 2. La Web Colombia como un objeto de estudio: Se fundamenta el proceso investigativo, se obtienen datos de la Web de Colombia, se aborda la forma en que se utilizó el crawler, como se configuró y con qué base se inició para lograr obtener información de la Web, se hace mención a la semilla y su composición, se compara con la cantidad de dominios que hacen parte de la Web en estudio y se establece una relación porcentual. Este apartado termina con una descripción de los datos encontrados y la forma en que se organizaron para su análisis.

El Capítulo 3. Caracterización de la Web Colombia, se tratan los resultados de la presente investigación, se describen los datos que se encontraron y los parámetros que se establecieron para el análisis de la misma, los cuales corresponden a: tamaño de las páginas, edad de las páginas, idiomas, tipos de páginas, lenguajes, documentos multimedia, archivos binarios y comprimidos en páginas, características de los sitios Web, nombres de sitios, números de páginas por sitio, profundidad de un sitio, cantidad de enlaces, entrantes, salientes e internos, características de los dominios, software utilizado como servidor, número de sitios por dominio, dominios genéricos y la estructura macroscópica de los mismos.

[3] TOLOSA, Gabriel, et. al. Caracterización del espacio Web de Argentina. [en línea]. < http://www.tyr.unlu.edu.ar/investigacion/web_Argentina_final.htm>. [citado el 27 de junio de 2016].

[4] CASTILLO, Carlos; GRAELLS, Eduardo y BAEZA-YATES, Ricardo. Características de la Web chilena. Centro de investigaciones de la Web. Volumen 1. Departamento de Ciencias de la Computación. Universidad de Chile. Chile. 1th edición. 2006. p. 58.

En los apartados finales se encontrará las conclusiones, recomendaciones, bibliografía y anexos, que son elementos importantes para la presente investigación y permitirán al lector conocer las fuentes de información utilizadas, algunos documentos adicionales importantes para el proceso.

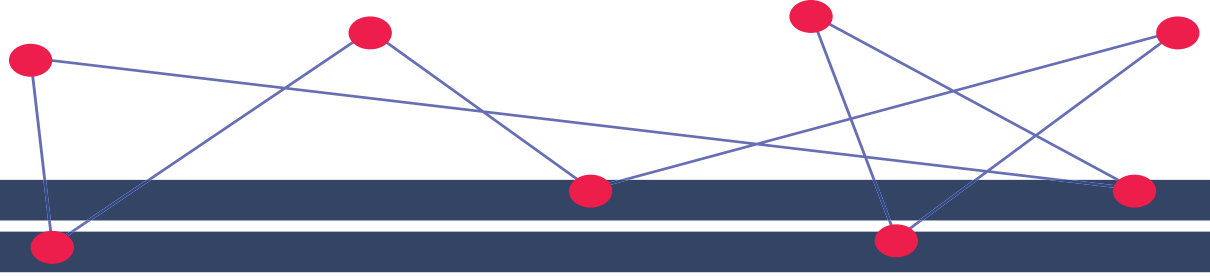
1. LA WEB

Este Capítulo empieza ...con la sección 1.1… que describe los principios fundamentales sobre la Web utilizados para el estudio, posteriormente ...la sección 1.2... ofrece una visión sobre los pilares que sustentan la arquitectura lógica. El...punto 1.3... esboza los conceptos sobre la filosofía de la Web orientados por Timothy Berners-Lee... la sección 1.4... desarrolla los conceptos de la web semántica. Posteriormente...en la sección 1.5...se profundiza en el análisis de la Web como un grafo dirigido. En...el parágrafo 1.6...se desarrolló el concepto de búsqueda mediante una arquitectura típica de una máquina web...el numeral 1.7... da a conocer los principales crawleis y su soporte...el ítem 1.8... presenta la arquitectura básica de un motor de búsqueda... el numeral 1.9... se desarrolló el concepto de los métodos de muestreo y finalmente en la sección…1.10... se describe el servicio DNS (*Sistema de Nombres de Dominio*), fundamentando la implementación de los NIC nacionales (*Network Information Center*).

1.1 Definición

Una necesidad básica humana es la de poder comunicarse, y en este sentido se ha pasado de conversaciones cara a cara a un gran número de herramientas actualmente. El desarrollo de la tecnología computacional ha sido clave en este proceso. Se ha pasado de grandes máquinas aisladas y escasas a unidades más funcionales, las cuales han dado origen a las redes de comunicaciones con el consecuente problema de cómo lograr compartir la información entre éstas máquinas. En este sentido y según Marcelo Arenas et. al[1]., a finales de 1.980 Timothy Berners-Lee licenciado en Física de la Universidad de Oxford, durante un trabajo de consultoría de software en el CERN (*Organización Europea para la Investigación Nuclear*) y en su tiempo libre, creó el primer programa tipo Web, con el fin de ayudarse a recordar las conexiones entre diversas personas, computadores y proyectos de laboratorio (*Figura 1*). En base a este concepto, pero con una idea mucho más amplia, la de poder compartir toda la información almacenada en distintos computadores en base a identificadores únicos, desarrolló los fundamentos que estructuran lo que hoy se conoce como la Web.

[1] ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 11.



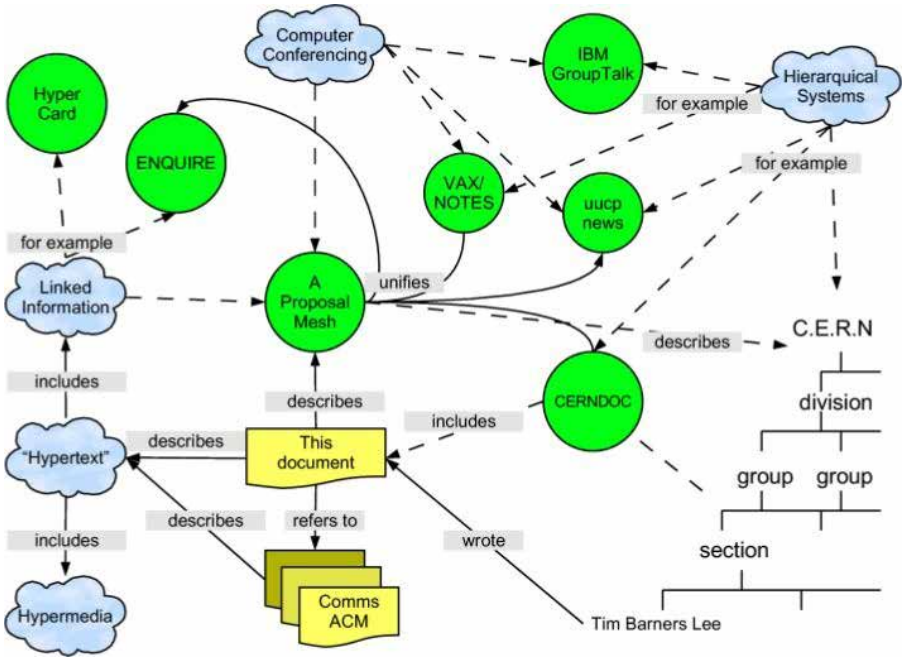
Para Federico Estupiñan y Xavier Molina[2] actualmente, la Web constituye un espacio público utilizado por millones de usuarios con diferentes objetivos. En sus inicios, se presentaba como un depósito distribuido que permitía compartir información pero con el avance tecnológico se ha convertido en la actualidad como un medio de publicación para diferentes usos como comercio, publicidad, educación, entretenimiento y contactos sociales, entre otros.

Internet y Web son términos que generalmente se utilizan para referirse a un sólo concepto, sin embargo se debe tener en cuenta que son elementos diferentes; Internet es un conjunto de redes que usan un conjunto de reglas comunes para ofrecer diversos servicios y la Web hace referencia a una aplicación o servicio que ha permitido que un sitio de Internet se establecieran páginas de información que pueden contener texto, imágenes, sonido, video y vínculos integrados a otras páginas, las cuales se encuentran en diferentes puntos de la red Internet acorde a lo expresado por Andrew Tanenbaum[3].

[2] ESTUPIÑAN, Mendoza Federico y MOLINA LARA, Xavier. Análisis de la web ecuatoriana mediante el uso de una herramienta de Web crawling. Escuela Politécnica Nacional, 2011. p.15.

[3] TANENBAUM, Andrew S. Redes de Computadoras. Prentice-Hall, Inc. Upper Saddle River. NJ. USA. 2003. p. 57.

Figura 1. Propuesta original de la Web en el CERN



Fuente: ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 12.

1.2 Bases lógicas de la web

Los tres pilares básicos sobre los cuales se sustenta la arquitectura lógica de la Web son los siguientes:

1. Identificadores únicos (URI): para poder referenciar y describir todos los objetos que hay en la Web, es necesario que éstos tengan su nombre propio, que en términos técnicos se llama identificador. En la Web estos nombres propios se llaman Identificadores Universales de Recursos. Una versión elemental de URI es la URL *(Localizador universal de recursos*), que corresponde a una dirección en la Web. Las direcciones representan una de las formas de identificar un objeto, pero hay que señalar que la noción de identificador es más amplia que la de dirección.

2. Lenguaje universal para describir HTML: para la comunicación universal es necesario un lenguaje único, entendible por todos. Timothy Berners-Lee diseñó el lenguaje HTML (*Hyper Text Markup Language*), que a más de su simplicidad de uso, suma una característica clave: el ser un lenguaje de hipertexto, es decir, que tiene una forma de anclar o redirigir al lector desde un punto cualquiera del texto a otro lugar. Técnicamente se les conoce como links o enlaces en la Web.

3. Protocolo de transmisión de datos HTTP: Desde un punto de vista más técnico, es necesario un protocolo que permita enviar y traer información en HTML desde un sitio a otro en la Web.

El protocolo HTTP (*Hyper Text Transfer Protocol*) tiene varias características distintivas que lo han hecho muy perdurable. HTTP es un protocolo de transmisión entre clientes y servidores. El cliente, que puede ser un browser o navegador, un agente, o cualquier herramienta. El servidor es el que almacena o crea recursos como archivos HTML, imágenes, etc. Entre ellos puede haber varios intermediarios, como proxy o gateways. A través de instrucciones simples, pero poderosas, el cliente indica al servidor qué acciones realizar para recibir o entregar datos[4].

1.3 Filosofía de la web

La Web fue creada con una cierta filosofía, una posición de principios frente a los desarrollos que se venían dando en materia de publicaciones, de desarrollo de software, de derechos de autor y de difusión. Esta filosofía puede resumirse en tres principios básicos: todos pueden publicar, todos pueden leer, nadie debe restringir.

¿Cómo lograr esto técnicamente? En esta dirección, se creó el Consorcio de la Web (*W3C*), una organización internacional que se propuso como sus dos objetivos primordiales el impulsar la interoperabilidad y evolutividad de la recientemente creada red universal de información. Para esto se comenzaron a generar estándares y protocolos. ¿Qué significan estos dos requerimientos en más detalle? En un famoso artículo, Explorando la Universalidad, Timothy Berners-Lee desglosaba sus aspectos básicos:

---

[4] ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edition, 2008. p. 13.

- Independencia de Dispositivo. La misma información debe ser accesible desde diversos dispositivos. Esto significa, por ejemplo, que la visualización debe tener estándares que permitan acceder a la información desde casi cualquier formato de pantalla y audio. Una de las bases para implementar esta desiderata es la separación de contenido y forma en la información.

- **Independencia de Software**. Hay muchos y diversos programas de software que se usan. Ninguno debe ser crítico para el funcionamiento de la Web. El desarrollo descentralizado del software ha sido clave para su crecimiento. Además, tema no menor, este postulado previene que la Web misma caiga bajo el control de una comunidad dada o algún gobierno usando el control del software.

- **Internacionalización**. Desde sus inicios, la Web no ha estado encargada a ningún país. Mediante el estándar de codificación de caracteres diseñado para facilitar el tratamiento informático, transmisión y visualización de textos de múltiples lenguajes y disciplinas técnicas denominada UNICODE esta barrera ha sido barrida.

- **Multimedia**. Los formatos disponibles para publicar deben estar abiertos a todas las facetas de la creatividad humana capaces de representar. En este sentido, soportar multimedia no representa sólo un par de avances tecnológicos, sino una filosofía de desarrollo de la Web.

- **Accesibilidad.** La gente difiere en múltiples cosas, en particular, en sus capacidades. La universalidad de la Web debe permitir que ella sea usada por la gente independientemente de sus discapacidades. De nuevo aquí la separación de contenido y forma de la información es un pilar básico.

- **Ritmo y razón.** Como afirma Tim Berners-Lee, la información varía desde un poema hasta una tabla en una base de datos. El balance entre procesamiento automático y humano debe estar presente. Por un lado, por las cantidades y tipo de información actualmente disponible es impensable que ésta sea procesada sólo por seres humanos: se necesitan agentes automáticos. Por otra parte, es absurdo pensar que en algún momento los humanos serán prescindibles en el desarrollo y enriquecimiento de la Web. Hay que buscar los justos términos para cada aplicación.

- **Calidad.** Las nociones de calidad son subjetivas e históricas. Por ello es impensable que algún día toda la información vaya a ser de calidad. Aquí hay otro compromiso, y es que la tecnología de la Web debe permitirnos navegar y vivir entre información con diferentes niveles de calidad.

- **Independencia de escala.** La armonía a gran escala supone armonía en sus componentes. La Web debe soportar grandes y pequeños grupos. Debe permitir que la privacidad de la información de individuos y grupos pueda ser negociada por ellos mismos, y permitir que cada grupo se sienta seguro en el control de su espacio. Hay que lograr un balance entre un gigante monolítico y una diversidad que pueda llevar al aislamiento completo de cada uno[5].

1.4 La Web Semántica

Uno de los problemas más importantes que aparece con la Web es el de determinar qué *significa* cada dato que está en la Web. Es prácticamente imposible para un usuario chileno entender una página en chino o tailandés. Y viceversa. El problema es aún más dramático: es muy difícil para un humano encontrar la información que necesita. Los buscadores funcionan de manera puramente *sintáctica*, es decir, no *entienden* las palabras. Por ejemplo, para encontrar todos los vuelos a Praga para mañana por la mañana, se obtendría unos resultados como los expresado en la *Figura 2*. *Buscador actual en la Web*, los cuales son inexactos ya que ofrece una información variada sobre Praga pero que no tiene nada que ver con lo que realmente el usuario busca. El paso siguiente por parte del usuario es realizar una búsqueda manual entre esas opciones que aparecen, con la consiguiente dificultad y pérdida de tiempo[6].

---

[5] ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edition, 2008. p. 14-16.

[6] WORLD WIDE WEB. Guías breves de tecnologías W3C. [en línea]. <http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica> [citado el 30 de junio de 2016].

Figura 2. Buscador actual en la Web

Fuente: WORLD WIDE WEB. Guías breves de tecnologías W3C. [en línea]. <http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica> [citado el 30 de junio de 2016].

Tradicionalmente eso era resuelto por *catalogadores*, personas especializadas que agregaban *metadatos* (*etiquetas que explicitan información*) a los libros: qué tema trata, dónde está ubicado, cuál es el autor, etc. Estos metadatos están accesibles en un catálogo en las bibliotecas. En la Web, no se tiene catálogo, ni menos catalogadores. Con los volúmenes de información que cada día crecen, es imposible que humanos se preocupen de clasificar la información. Además, porque el modelo de la Web es distribuido, quienes publican tienen diversas visiones sobre cómo clasificar sus objetos.

Para los profesionales de la información, el principal desafío hoy es cómo manejar esta extraordinaria cantidad de datos que crece día a día. Estamos comenzando a ver los problemas: los motores de búsqueda a menudo no contestan lo que buscamos; hay dificultades para filtrar la información; la heterogeneidad de los datos y los contenidos; desde el punto de vista de quien publica, se ha convertido en un problema hacer visible lo visible, tanto en formato como en contenido. Ha habido avances en los niveles estructurales y sintácticos con el estándar XML y sus tecnologías aledañas.

Desafortunadamente, al nivel del significado (*semántica*) aún estamos muy por debajo de las necesidades. Estamos lejos de responder preguntas como "todos los museos que exhiban trabajos de Guayasamín" o "¿Cuál es la biblioteca que tiene la mejor colección de los escritos de Gandhi?" o "¿Cuál es la compañía que ofrece el mejor mapa de Isla de Pascua desde el punto de vista precio/resolución?" Un motor de búsqueda estándar *(como Google, Yahoo!, etc.*) no puede responder tales consultas. Pero tampoco ningún agente las podría responder hoy en día. Sin embargo, la información está allí: hay que relacionarla y agregarla. La limitación obedece a la falta de habilidad de las máquinas para entender el significado y las relaciones entre las partes de información que recolectan. Hoy en día los humanos agregan el contexto, interpretan y dan sentido a la información que existe en la Web. En otra dirección, otro ejemplo de estas limitaciones es la dificultad para diseñar e implementar una tarea tan natural como organizar todos los recursos educacionales de un país, de tal forma que resulte sencillo para cada estudiante y profesor el publicar y obtener la información que requieran. Se necesitan vocabularios comunes, descripción precisa de los datos expuestos, publicación distribuida, búsquedas automatizadas. En una frase: debido a las enormes dimensiones, la Web se ha convertido en una torre de Babel no sólo al nivel del lenguaje natural, sino esencialmente al nivel del significado, contradiciendo las ideas por las cuales fue creada. ¿La solución? Pavimentar el camino para la construcción de agentes de software que puedan procesar información de la Web por nosotros. La noción de *Web Semántica* es transformar la Web actual de tal forma que la información y los servicios sean entendibles y usables tanto por computadores como por humanos. La Web Semántica creará el ambiente necesario donde los agentes de software puedan rápidamente realizar tareas sofisticadas y ayudar a los humanos a encontrar, entender, integrar, y usar la información en la Web.

La *Figura 3* muestra los resultados obtenidos a través de un buscador semántico. Estos resultados ofrecen al usuario la información exacta que estaba buscando. La ubicación geográfica desde la que el usuario envía su pregunta es detectada de forma automática sin necesidad de especificar el punto de partida, elementos de la oración como "mañana" adquirirían significado, convirtiéndose en un día concreto calculado en función de un "hoy". Algo semejante ocurriría con el segundo "mañana", que sería interpretado como un momento determinado del día. Todo ello a través de una Web en la que los datos pasan a ser información llena de significado. El resultado final sería la obtención de forma rápida y sencilla de todos los vuelos a Praga para mañana por la mañana[7].

[7] WORLD WIDE WEB. Guías breves de tecnologías W3C. [en línea]. <http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica> [citado el 30 de junio de 2016].

Figura 3. Buscador semántico

Fuente: WORLD WIDE WEB. Guías breves de tecnologías W3C. [en línea]. <http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica> [citado el 30 de junio de 2016].

1.4.1 Metadatos y RDF

La característica distintiva de la Web Semántica será un lenguaje estándar de metadatos y ontologías, que permitirán que agentes de software encuentren el significado de la información en páginas Web, siguiendo enlaces a las definiciones de términos claves y reglas para razonar acerca de ellas lógicamente. Los metadatos son datos descriptivos acerca de un objeto o recurso, sea éste físico o electrónico. Las ontologías son especificaciones formales de vocabulario y conceptos compartidos para un dominio.

Aunque el concepto de metadatos es relativamente nuevo, los conceptos subyacentes han estado rondando desde que se organizaron grandes colecciones de información. En áreas tales como catalogación en bibliotecas y museos han sido usados por décadas. Una manera útil de pensar acerca de los metadatos es "la suma total de lo que no puede decir acerca de cualquier objeto de información a cualquier nivel de agregación". Hay muchos tipos de metadatos, y los usos más comunes se refieren a documentación de copy rights y accesos legales, versionamiento, ubicación de información, indización, descripción de condiciones físicas de recursos, documentación de software, autentificación, etc.

En la Web, los metadatos también han jugado un rol importante en áreas como catálogos de propósito general (*como por ejemplo Wikipedia*), sin-

dicación y rating (*Rich Site Summary RSS, Platform for Internet Content PICS*), colecciones personales (*música, fotos*), privacidad, etc. Y los más populares hoy son simplemente tags, es decir, etiquetas; un lenguaje que no tiene verbos ni adjetivos. Simplemente nombres. Todos estos metadatos son sectoriales y usan una diversidad de modelos y lenguajes.

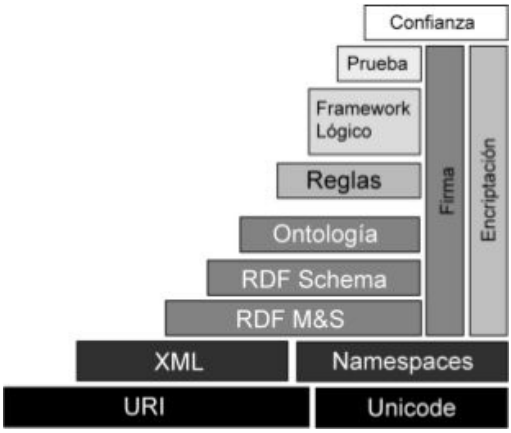
Por el contrario, se necesita un lenguaje de significados (*de metadatos*) universal. Este es RDF (*Resource Description Framework*), que es un lenguaje diseñado para soportar la Web Semántica, de la misma manera que HTML es el lenguaje que ayudó a iniciar la Web. El modelo de RDF es simple: el universo a modelar (*la Web*) es un conjunto de recursos (*esencialmente todo puede tener una URL*); el lenguaje para describirlo es un conjunto de propiedades (*técnicamente predicados binarios*); las descripciones son oraciones similares en estructura al modelo sujeto-predicado-objeto, donde el predicado y el objeto son recursos o cadenas de caracteres (*Figura 4*). Así, por ejemplo, uno puede afirmar "El creador de http://www.picarte.cl es Claudio Gutiérrez". El vocabulario de las propiedades para este lenguaje puede ser definido siguiendo las líneas dadas en los esquemas RDF (*RDF Schema*), y básicamente son codificaciones de ontologías a diferentes niveles[8].

1.4.2 El Futuro de la Web

No es fácil predecir los desarrollos futuros de la Web. El proyecto inicial de Tim Berners-Lee incluía el desarrollo de capas sucesivas para permitir el intercambio global de información y conocimiento. Luego de la estructura básica que conocemos, vendrá una capa de semántica, de metadatos. Esta capa permitiría procesar la información semi-automáticamente, es decir, permitiría a agentes de software procesar la información en paralelo a los humanos. *(Nótese que la Web actual está hecha casi en su totalidad para que seres humanos la naveguen.)*

[8] WORLD WIDE WEB. Guías breves de tecnologías W3C. [en línea]. <http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica> [citado el 30 de junio de 2016].

Figura 4. La torre de la Web Semántica



Fuente: ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 20.

La Web por supuesto ha evolucionado en miles de direcciones, muchas no previstas, como redes sociales, blogs, etc. Muchos han llamado al conjunto de estos desarrollos "novedosos" no previstos Web 2.0.

El futuro está abierto. Hoy en día no es posible predecir los usos futuros de la Web, y aquí ya entramos al campo de la ciencia ficción. [9]

1.5 Anatomía

Actualmente no existe en internet un sistema de control central que permita establecer aspectos como la estructura y el tamaño de la Web, pese a eso hay una serie de compañías que pueden dar una aproximación, worldwidewebsize estima que el tamaño de la Web es de 4.02 billones de páginas indexadas[10], por lo cual es necesario generar sistemas de búsqueda que permitan aprovechar esta gran base de datos compuesta por imágenes, texto, audio y video.

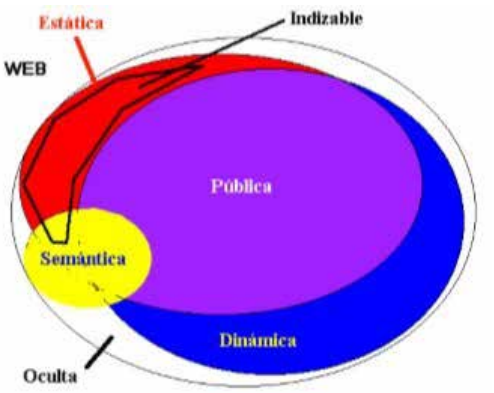
---

[9] ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 16-21

[10] THE SIZE OF THE WORLD WIDE WEB. [en línea]. <http://www.worldwidewebsize.com/>. [citado el 30 de junio de 2016]

La Web dispone de estos elementos mediante el uso de las páginas, las que pueden ser estáticas y dinámicas, públicas y privadas, con o sin metadatos, como se muestra la *Figura 5*. Las páginas estáticas son aquellas que existen todo el tiempo en un archivo en algún servidor Web. Las páginas dinámicas son aquellas que se crean cuando una persona interactúa con un servidor Web, por ejemplo la respuesta a una consulta en un buscador o el resultado de rellenar un formulario en un sitio de comercio electrónico. Actualmente, la mayor parte de la Web es dinámica, y como en algunos sitios se puede generar un número no acotado de páginas dinámicas (por ejemplo, un calendario), la Web que podemos crear es infinita. Las páginas públicas son las que todas las personas pueden ver y las privadas son las que están protegidas por una clave o se encuentran dentro de una Intranet. Como cada persona tiene acceso a distintas páginas privadas, la Web pública depende del observador. En particular cada buscador refleja una Web pública distinta. Algunos sitios tienen información semántica que ayuda a los buscadores y se estima que un 5% de ellos tiene información fidedigna. Sin embargo, más son los sitios que tienen información falsa, lo que se llama spam de Web[11].

Figura 5. Principales características de la Web



Fuente: ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 24.

[11] ARENAS, MARCELO e.t.al.Cómo funciona la web volumen 1. Universidad de Chile , Santiago de Chile, 1th edición, 2008. p. 23-35

"La Web es más que un simple conjunto de documentos en distintos servidores, ya que existen relaciones de información entre los documentos mediante los enlaces que establecen entre ellos. Debido a esto se plantea que sigue un modelo de grafo dirigido, en el que cada página es un nodo y cada arco representa un enlace entre dos páginas"[12].

En él existe un componente fuertemente conexo, el cual es un subconjunto de los nodos del grafo donde existe un camino entre cualquier par de ellos. Los componentes fuertemente conexos que poseen más de un sitio no son demasiados y el de mayor tamaño se llama principal, core o MAIN. Según Baeza et. al.[13] se puede realizar una clasificación de los sitios de un espacio Web de acuerdo a su relación con el componente fuertemente conexo principal (Figura 6).

Todo sitio nuevo debe entrar a la estructura de la Web, estos son difíciles de encontrar sin campañas de publicidad, correo electrónico o a través de comunicación verbal. Lo mismo pasará con los buscadores como Google o Yahoo!, que usan los enlaces a un sitio para evaluar su importancia[14]. De tal forma que generalmente un sitio comenzará en ISLAS o IN (*sitios que llegan a MAIN, pero desde MAIN no se puede llegar a ellos*). Luego, si es conocido, pasa al centro de la web o MAIN (componente fuertemente conexo). Si luego decide no apuntar a un sitio importante o no es actualizado pasa a la derecha u OUT (*sitios a los que se llega desde MAIN, pero no se puede retornar*), o peor aún, se convierte nuevamente en ISLAS.

Dentro de MAIN, pueden estar ubicados como MAIN-MAIN que son sitios relacionados directamente con IN y con OUT, MAIN-IN sitios relacionados directamente con IN, pero no con OUT, MAIN-OUT sitios relacionados directamente con OUT pero no con IN y los MAIN-NORMAL que son los sitios en MAIN que no corresponden a ninguna de las categorías vistas.

Los TUNNEL son los caminos de los sitios que llegan de IN a OUT sin pasar por el MAIN y los TENTACLE son los sitios a los que se llega de IN o van a OUT, y no están en MAIN ni en TUNNEL.

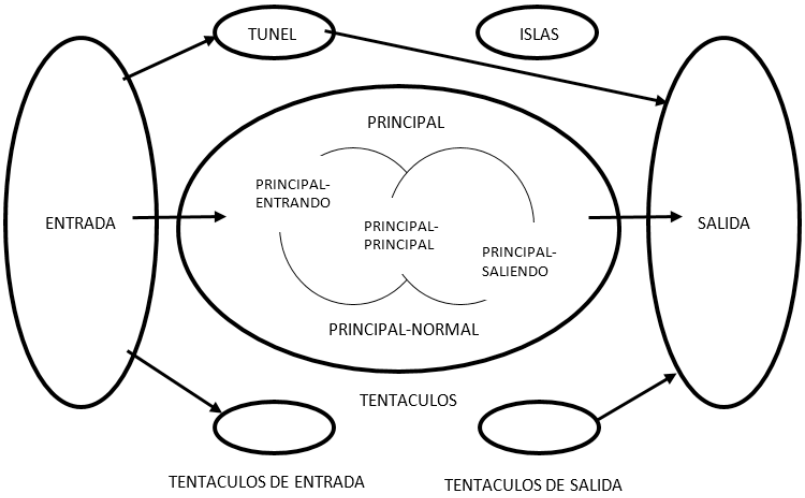
---

[12] CASTILLO, Carlos; GRAELLS, Eduardo y BAEZA-YATES, Ricardo. Características de la Web chilena. Centro de investigaciones de la Web. Volumen 1. Departamento de Ciencias de la Computación. Universidad de Chile. Chile. 1th edición. 2006. p. 5.

[13] Ibíd., p. 34.

[14] ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 37.

Figura 6. Estructura macroscópica de la Web



Fuente: TOLOSA, Gabriel, et. al. Caracterización del espacio Web de Argentina. [en línea]. < http://www.tyr.unlu.edu.ar/investigacion/web_Argentina_final.htm>. [citado el 27 de junio de 2016].

Esto presenta muchas ventajas, tanto para los usuarios, a la hora de buscar información, como para los programas que recorren la Web, a la hora de buscar contenido para recolectar (probablemente para un motor de búsqueda).

1.6 Búsqueda

Para operar este milagro no basta con Internet. Ni siquiera basta con la Web. El ingrediente imprescindible que se necesita son los buscadores o máquinas de búsqueda. Estos buscadores, cuyos representantes más conocidos son probablemente Google, Yahoo! y Microsoft MSN, son los que conocen en qué páginas de la Web aparecen qué palabras (y saben bastante más). Sin un buscador, deberíamos conocer las direcciones Web de todos los sitios de bibliotecas, o de turismo, o de cualquier tema que nos pudiera interesar, y los que no conociéramos sería como si no existieran. En un sentido muy real, los buscadores conectan la Web, pues existen grandes porciones de la Web a las que no se puede llegar navegando desde otra parte, a menos que se use un

buscador. No es entonces sorprendente que casi un tercio del tiempo que los usuarios pasan en Internet lo dediquen a hacer búsquedas.

La arquitectura típica de una máquina de búsqueda se observa en la *Figura 7*. En el crawling se recolectan páginas de la Web, ya sea nuevas o actualizadas. El proceso de indexamiento es el que extrae los enlaces que parten de las páginas leídas y realimenta el crawling con nuevas direcciones para visitar, mientras que almacena en el índice la información para qué palabras aparecen en qué páginas, junto con una estimación de la importancia de tales ocurrencias. La búsqueda usa el índice para responder una consulta, y luego presenta la información al usuario para que éste navegue por ella[15].

1.7 Crawlers

"Las herramientas que se usan para analizar estos procesos se denominan crawlers, que son programas o scripts automatizados. Su función es ir recorriendo todos los dominios a partir de un punto inicial prefijado, descargando el contenido de los sitios atravesados"[16]. Se debe tener una buena estrategia de rastreo, pero también necesita de una gran optimización en su arquitectura. Shkapenyuk et. al. señala que:

"Si bien es bastante fácil de construir un crawler lento que descargue algunas páginas por segundo durante un corto periodo de tiempo, la construcción de un sistema de alto rendimiento que puede descargar cientos de millones de páginas durante varias semanas presenta una serie de desafíos en el sistema, I/O, eficiencia de la red, solidez y la capacidad de gestión"[17].

---

[15] ARENAS, Marcelo et. al. Cómo funciona la Web, volumen 1. Universidad de Chile, Santiago de Chile, 1th edición, 2008. p. 52.

[16] GASCON, Álvaro; DE LA PUENTE, Marín y RODRIGUEZ APARICIO, Miguel María. Clasificación jerárquica de contenidos web. España, Universidad Carlos III. p. 2.

[17] SHKAPENYUK, Vladislav y SUEL, Torsten. Design and implementation of a high-performance distributed web crawler. In In Proc. of the Int. Conf. on Data Engineering. 2002. p. 357–368.

Caracterización de la Web Colombia mediante la herramienta WIRE

EDITORIAL
Institución Universitaria CESMAG