

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

SEMANTIC WEB VS. DEEP WEB: MITOS Y REALIDADES

Andrés Felipe Jiménez Vélez

Diego Alejandro Villada Vásquez

FACULTAD DE INGENIERÍAS

Ingeniería de Sistemas

Director
Prof. Edgar Serna M.

INSTITUTO TECNOLÓGICO METROPOLITANO

Febrero 2016

RESUMEN

Cuando la mayoría de personas piensa en la Web hace referencia a la web visible, es decir, las páginas que pueden acceder directamente a través de enlaces o por medio de los motores de búsqueda comunes. Sin embargo esta web representa solamente el 4% por ciento de todo el contenido en Internet. La Web se ha convertido rápidamente en una entidad titánica y compleja y se estima que cada año, a partir de 2015, se añadirá un zetabyte o más de contenido. Navegar por este mar de información representa cada día un desafío, sobre todo cuando gran parte de ese contenido no se accede fácilmente a través de los tradicionales motores de búsqueda. En términos generales a la web que se puede acceder de forma directa se le conoce como Semantic Web y a la oculta Deep Web. Debido a que se ha tejido todo un universo de mitos y creencias alrededor de cada una, el objetivo de esta investigación es describir la realidad de esta situación y de divulgar los resultados como fuente de información para las diferentes comunidades.

Palabras clave: Web semántica, web profunda, ciberseguridad, gestión del conocimiento.

RECONOCIMIENTOS

Agradecemos profundamente a nuestras familias por su entrega incondicional, paciencia, esfuerzo y perseverancia; fueron pilares fundamentales en el desarrollo personal y profesional, que en conjunto con todos los logros alcanzados en el transcurrir de nuestras vidas académicas, han permitido construir un futuro promisorio y acorde con las metas trazadas en el inicio de nuestras carreras.

También agradecemos a la institución univertaria ITM que nos acogió desde el inicio, proporcionando una planta física y docente de alta calidad; sin su acompañamiento y estímulo personal y profesional no estaríamos contemplando estos buenos resultados.

Un agradecimiento especial a nuestro asesor de trabajo de grado el docente Edgar Serna, su acompañamiento e instrucción permitió llevar a buen término este gran aporte académico, que esperamos sea de gran utilidad para futuros estudiantes interesados en el área.

ACRÓNIMOS

CRAWLER Programa que navega sistemáticamente por la World Wide Web con el fin de crear un índice de datos. (Conocido también como araña). Va saltando de una página a otra según los enlaces que vaya encontrando en cada una.

METADATO conjunto de datos que describen otros datos o el contenido de información de un objeto denominado recurso.

ONTOLOGIA Modelo de datos que representa conocimiento, como un conjunto de conceptos dentro de un dominio y las relaciones entre esos conceptos.

RDF Modelo de datos para los recursos y las relaciones que se puedan establecer entre ellos, adicionalmente aporta la semántica para establecer jerarquías de generalización entre las propiedades y clases.

RDFS Vocabulario que permite describir las propiedades y las clases de los recursos

SEMANTIC WEB Red informática donde residen un conjunto de recursos publicados de acuerdo con un conjunto de estándares para que las máquinas los puedan interpretar. Estos manejan etiquetado que describe tanto el formato del documento como el contenido (etiquetas describen las entidades del documento: personas, organizaciones)

TOR (The onion router): "Software libre y red abierta, que ayuda a defenderse en contra del análisis de tráfico, una manera de defenderse de la vigilancia en la red que amenaza la privacidad y libertad personal, actividades y relaciones de negocios, y seguridad de los estados."

WEB SUPERFICIAL (Surface Web): Es la parte de internet donde se encuentran la mayoría de sitios Web accesibles a través de los motores de búsqueda convencionales por estar indexados.

WEB PROFUNDA (Deep Web): Es la parte de internet en la cual las arañas (crawlers) de los buscadores no tienen acceso debido a varios factores presentes en las páginas Web, los más comunes son:

La página Web tiene un archivo robots.txt con el código Disallow: * que impide la entrada de cualquier araña que respete las normas de este archivo.

La página Web está programada en Flash, cosa que es imposible de leer por los buscadores.

	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

La página Web está protegida con contraseña y esta es desconocida por los robots de indexación.

La página Web se genera de forma dinámica por lo tanto cuando una araña accede, no encuentra ningún tipo de información que indexar.

La página Web no contiene enlaces entrantes que el robot pueda seguir para llegar a ella desde otros sitios (Victoria Coll, 2015)

XMLS Es el lenguaje que permite definir la estructura de documentos XML

TABLA DE CONTENIDO

1. INTRODUCCIÓN	4
2. MARCO TEÓRICO.....	8
3. METODOLOGÍA	10
4. RESULTADOS Y DISCUSIÓN	11
5. CONCLUSIONES	24
REFERENCIAS	26
APÉNDICE (Artículo científico)	28

1. INTRODUCCION

Cuando se piensa en la web, la mayoría de personas hace referencia a la web visible, es decir, las páginas que se pueden acceder directamente a través de su enlace o desde los motores de búsqueda. Sin embargo, esta imagen de la web solamente representa el 4% de todo el contenido en internet. La web se ha convertido rápidamente en una entidad compleja y se estima que cada año se añade un *zetabyte*, o más, de contenido. Navegar en esta cantidad de información representa un desafío, especialmente porque a través de los motores de búsqueda tradicionales no se accede fácilmente a ella. En términos generales, a la web que se puede acceder de esta forma se le conoce como *Semantic Web*, y a la parte *oculta* o *invisible*, que se accede solamente a través de herramientas especializadas o páginas con control de acceso, se le conoce como *Deep Web*. Esta última recibe su nombre debido a la incapacidad de los navegadores tradicionales para acceder a la información no-catalogada o indexada que se encuentra en ella. Esto se debe a que los motores de búsqueda todavía no incorporan los nuevos estándares de indexación y no tienen la capacidad de acceder a páginas dinámicas. Otras razones van desde el bloqueo impuesto desde la estructura de sus sitios, porque están protegidos por contraseña de seguridad mediante variables *no follow* para evitar el seguimiento de la huella digital, o simplemente porque contienen formatos que no pueden ser interpretados por los robots *crawlers*, encargados buscar e indexar la información. En estos casos, el motor de búsqueda no encuentra el código que leen los robots en el contenido asociado para realizar referencia estática, o simplemente no pueden acceder a las bases de datos protegidas, por lo que pasan de largo ante estos sitios (Shestakov, 2008).

Ian Clarke (1999) propuso un sistema de información de almacenamiento y recuperación distribuido y descentralizado, o como algunos podrían afirmar, una nueva y revolucionaria forma de utilizar Internet sin ser detectado. Es decir, con este software, cualquier persona podría conversar en línea, o leer o crear una página web, o compartir archivos con anonimato casi total. Su objetivo era evitar los estrictos controles que

suponía la red en ese entonces, porque para él Internet podía ser controlado con mayor rapidez, de forma más completa y más barata que con los sistemas de comunicación de la época, y su software pionero tenía la intención de cambiar eso.

La red moderna a menudo es considerada como un milagro tecnológico, en el que aparentemente todo funciona sin contratiempos. Muchos usuarios piensan que cuando realizan una búsqueda a través de un motor cualquiera, van a tener acceso a toda la información en todas las páginas web. La realidad es que ese buscador solamente accede a una fracción muy pequeña de la verdadera profundidad de la web, y trae a la superficie pequeños fragmentos de la información disponible. Aunque no se tiene forma de estimar cuán grande y profunda es la web, sí se tienen certeza de que el área en la que navega la mayoría de internautas es una fracción muy pequeña de ella. Al resto se le ha denominado de diversas formas: *darkweb*, *deep web*, *surface web*, *invisible web*, *dark address space*, *murky address space*, *dirty address space*, y esas metáforas sólo hacen que internet se convierta en más insondable y misteriosa cada vez. La razón es que no todos estos eufemismos significan lo mismo: mientras que *darknet* es una red en línea, como Freenet, que se oculta a los no-usuarios, con todo el potencial que implica en cuanto comportamiento transgresor, gran parte de *deep web*, atemorizante como suena, consiste del consumo e investigación de datos sin complicaciones, pero que está más allá del alcance de los motores de búsqueda tradicionales. Por otro lado, *dark address space* se refiere a menudo a las direcciones de Internet que, por razones puramente técnicas, simplemente han dejado de funcionar.

Pero, en cierto sentido, todos forman parte de una misma imagen, en la que más allá de los confines de la vida en línea de la mayoría de personas, existe por ahí otra inmensa internet, utilizada por millones, en gran parte ignorada por los medios pero bien entendida solamente por unos pocos científicos computacionales. A finales de los años 90, Bergman (2001) realizó una investigación para tratar de medir el tamaño de internet. Partió de la hipótesis de que, conociendo la web de acceso cotidiano, era probable que la otra web fuera dos o tres veces más grande. Pero, luego de más de diez años de trabajo, la inmensidad de la misma todavía lo deja sin aliento, y sigue girando sobre supuestos y sin descubrir lo que se propuso entonces. Solamente se ha acercado con algunas cifras y suposiciones, como que actualmente esta web es 400 a 550 veces

más grande que la Semantic Web, que es la de más rápido crecimiento en nueva información, que el valor de su contenido es inconmensurable y que los buscadores tradicionales solamente tienen acceso al 0,03% del total de las páginas disponibles en internet.

Pero su tamaño no es el único problema, también hay que enfrentarse a los sitios bloqueados. Muchas veces la motivación es comercial, porque han invertido mucho tiempo y dinero como para dejar la información libre, otras veces es por intereses delictivos, cuando el contenido del sitio está dirigido a cometer actos ilícitos y no le conviene salir a la luz pública. De hecho, internet siempre ha estado impulsado por el deseo de mantener el secreto como una idea de transparencia. La red es una creación conjunta del Departamento de Defensa de los Estados Unidos y WELL, una comunidad de contracultura americana y una de las primeras y más influyentes en la web. Ambos grupos tenían razones para construir un entorno oculto o semi-oculto en línea, lo mismo que otro abierto. Por otro lado, todavía existen rincones secretos de internet donde perdura esta extraña alianza entre liberales y militares. Pero, para la mayoría de personas y empresas, la web es un universo en el que pueden comprar, vender, intercambiar y conocer sitios y personas en todo el mundo. Por todo esto, parece probable que Internet permanecerá un tiempo más en su fase de fiebre del oro, por lo que en las grietas y esquinas de sus estructuras, seguirán apareciendo sitios oscuros y otros entornos en línea privados.

Estos y otros factores son los que generan mitos y creencias alrededor de ambas plataformas, y los usuarios han creado su propia imagen de ellas. Ya sea porque han tenido experiencias poco agradables en la navegación, o porque los comentarios son malintencionados, o porque se exagera el escenario que han vivido, en la red se encuentra todo tipo de referencias a la Semantic Web y a la Deep Web. Debido a que alrededor de ambos entornos de navegación se ha tejido un universo de mitos y creencias, es necesario aclararlos de la mejor forma.

El objetivo de este artículo es presentar un análisis a la realidad Semantic Web vs Deep Web, desde un punto de vista objetivo e imparcial. La idea es ofrecer pautas para que las personas comprendan las ventajas y desventajas de cada una, y presentarles los criterios de seguridad y de riesgo a los que se exponen cada día en la red. Este trabajo

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

se encuentra estructurado de la siguiente forma: en la primera parte se detallan los trabajos relacionados con lo que se presenta, en la segunda se hace el análisis teórico a la realidad de Semantic Web y Deep Web y en la tercera se presentan las conclusiones respectivas.

2. MARCO TEORICO

Ian Clarke (1999) propuso un sistema de información de almacenamiento y recuperación distribuido y descentralizado, o como alguien con menos conocimiento podría decir, una nueva forma revolucionaria para utilizar Internet sin ser detectado. Es decir, con este software cualquier persona podría conversar en línea, leer o crear una página web o compartir archivos, casi con anonimato total. Su objetivo era evitar los estrictos controles que suponía la red, porque para él Internet podía ser controlado con mayor rapidez, de forma más completa y más barata que con los existentes sistemas de comunicación, y su software pionero tenía la intención de cambiar eso.

El internet moderno a menudo se considera como un milagro de la apertura, por su alcance es global, su flanqueo de censores y los motores de búsqueda, una cuestión que aparentemente todos ven. Por eso es que la mayoría de usuarios piensan que cuando hacen una búsqueda van a obtener toda la información en todas las páginas web. La realidad es que solamente tienen acceso a la pequeña fracción que los motores de búsqueda pueden acceder desde la Semantic Web (Sherman y Pric, 2003).

Deep Web, Dark Web y Surface Web son metáforas que solamente hacen que internet sea percibida más insondable y misteriosa. Otros términos circulan entre los que saben: Dark Net, Invisible Web, Dark Address Space, Murky Address Space o Dirty Address Space. Pero no todos significan lo mismo. Mientras que Dark Net es una red en línea que se oculta a los no-usuarios, con todo el potencial de la conducta transgresora que implica, gran parte de Deep Web, espeluznante como suena, se compone de datos de consumo corriente y de investigación que están más allá del alcance de los motores de búsqueda tradicionales. Dark Address Space a menudo se refiere a direcciones de Internet que, por razones puramente técnicas, simplemente han dejado de trabajar. Sin embargo, en cierto sentido, todos estos términos hacen parte de la misma imagen: más allá de los confines de la vida en línea de la mayoría de navegantes, existe una vasta internet utilizada por millones, pero en gran medida ignorada por los medios de comunicación aunque bien mantenida por unos pocos científicos computacionales. ¿Cómo se creó? ¿Qué ocurre exactamente en ella? ¿Qué representa el futuro o el pasado

de la vida en línea? Mientras que Semantic Web es la superficie de la red, en la que navega la mayoría de usuarios, solamente representa una porción de la cantidad de la información que se encuentra en Deep Web (Bergman, 2001). Para este autor, la diferencia es profunda y actualmente puede llegar a ser 500 o 600 veces más grande que la primera. Deep Web es la que tiene más rápido crecimiento en nueva información, el valor de su contenido es inconmensurable y las búsquedas en Internet solamente cubren el 0.03% del total de páginas disponibles.

Internet siempre ha sido impulsado tanto por el deseo de mantener el secreto como por una voluntad de transparencia. La red fue la creación conjunta del departamento de defensa de Estados Unidos y de la contracultura americana, y ambos grupos tenían razones para construir entornos ocultos o semi-ocultos así como también los abiertos. Por esto es que el cifrado fuerte fue desarrollado en paralelo con internet (O'Brien, 2005). Actualmente todavía persiste esa tendencia, con la complicación de que en Semantic Web es donde se libran las peores batallas.

Semantic Web abarca una serie de estándares que a menudo sirven de base para definir ontologías con referencia al contexto, y el lenguaje XML se ha convertido en un medio ampliamente aceptado de comunicación en ella, porque puede agregar semántica a los datos almacenados. El proyecto Semantic Web prevé que las personas publiquen información semántica en un formalismo procesable en computador, que permite que sea interconectada globalmente. Esos formalismos están basados en XML, pero van más allá de ella y especifican las relaciones semánticas entre entidades e incluso sus limitaciones lógicas. La colección de conocimiento del mundo en estos formalismos se conoce comúnmente como una ontología (Chang et al., 2006).

Para He et al. (2007), gran parte de la información presente en la World Wide Web se oculta a los motores de búsqueda porque no es accesible a través de hipervínculos, pero se encuentra en las bases de datos que pueden ser consultadas a través de otras formas. Los volúmenes de información son ampliamente diferentes y es indudable que aquí es de mejor calidad. Pero debido a los falsos rumores y a que también es cierto que la utilizan personas y organizaciones con intenciones oscuras, se ha mantenido al margen con relativamente poca publicidad (Staab y Studer, 2008).

3. METODOLOGIA

La metodología para realizar este proyecto consiste en consultar a los autores que han experimentado la navegación en ambos desarrollos y que han publicado sus resultados con alto rigor de credibilidad. Para lograrlo se plantean las siguientes fases:

1. Encontrar o construir una definición amplia y estructurada de Semantic Web y Deep Web.
2. Definir criterios para evaluar el rigor de los resultados publicados.
3. Consultar y analizar las fuentes directas e indirectas.
4. Estructurar y presentar los resultados en un artículo científico.

CRONOGRAMA DE ACTIVIDADES

Actividad	Meses					
	1	2	3	4	5	6
1. Encontrar o construir una definición amplia y estructurada de Semantic Web y Deep Web.	X					
2. Definir criterios para evaluar el rigor de los resultados publicados.	X					
3. Consultar y analizar las fuentes directas e indirectas.		X	X			
4. Estructurar y presentar los resultados en un artículo científico.				X	X	X

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

4. RESULTADOS Y DISCUSION

ANÁLISIS A LA REALIDAD DE LA WEB

World Wide Web es hoy por hoy el repositorio más grande y completo que haya existido en la historia de la información. De acuerdo con la Unión Internacional de Telecomunicaciones, a finales de 2014 se estimaba que internet tenía 3.000 millones de usuarios, una cantidad que crece cada día y para la que es necesario ofrecer servicios adaptando las tecnologías actuales y mejorando los tiempos y la calidad de las respuestas. Para lograrlo, es necesario intervenir los modelos de desarrollo y mecanismos de escaneo e indexación que emplean los navegadores, para ofrecer con éxito las respuestas a las consultas de los usuarios. De lo contrario, la información contenida en los miles de sitios alojados en todo el mundo se podría desperdiciar (Calaresu y Shiri, 2015). El problema aquí es que el volumen de información en la Web crece cada día, pero las herramientas para consultarla no se desarrollan al mismo ritmo. Esto crea una especie de brecha entre la disponible y la que se puede acceder. Para la mayoría de usuarios esto no es representativo, porque sus necesidades de búsqueda se satisfacen fácilmente. Pero a los científicos y gobiernos se les dificulta realizar la mayoría de sus consultas de gestión del conocimiento, porque no es fácil acceder a la información con las herramientas actuales (Fensel et al., 2003).

Parte de toda la información documentada del planeta está disponible a través de los motores de búsqueda convencionales, que muchos denominan Web superficial. Sin embargo, existe un segmento mucho mayor que no se encuentra indexado por ellos y no es de fácil acceso, a esa Web se conoce ampliamente como Web profunda. Se estima que esta última contiene más de 2.5 billones de documentos, y que crecen a un ritmo de 7.5 millones por día. Además, debido a que la información en la red contiene enlaces de relación, los documentos que se encuentran dentro de otros documentos son más propensos a ser indexados en otros nuevos, y los buscadores tradicionales no los logran identificar. Para superar estos problemas, se ha desarrollado una nueva generación de motores de búsqueda en la cual se reemplaza el *link following* al azar, enfocándose en el rastreo y la indexación dirigida al tráfico de las páginas. Este enfoque genera mejores

resultados en consultas simples, aunque aumenta la tendencia a pasar por alto información poco referenciada (Bergman, 2001).

Para comprender el proceso de ubicación de la información en internet, Marc Victoria (2015) detalla los niveles de clasificación de los sitios en la red, tal como se observa en la Figura 1. Al nivel 0 tienen acceso todos los usuarios de la web a través de motores de búsqueda convencionales, y las consultas muestran las páginas que fueron indexadas por el buscador. Este nivel cubre la porción más pequeña de la información documentada en internet. En los niveles 1 y 2 se puede consultar algunos sitios desde la Web superficial, y a partir del nivel 3 se ingresa al mundo de Deep Web, que es accesible a través de tecnologías o sitios especializados y en muchos casos solamente a través del pago del servicio.



Figura 1. Niveles de clasificación de las páginas web (Victoria, 2015)

Semantic Web

Cuando se realiza una búsqueda en la web, los navegadores realizan esta actividad en bases de datos especializadas que constantemente son actualizadas por *web crawlers*, una especie de robot que ejecuta consultas constantes entre enlaces de hipertexto para añadir la información que encuentran a sus bases de datos. Con esta operación proactiva agilizan la entrega de resultados y mantienen actualizadas las diferentes

entradas, ya sea eliminándolas cuando el sitio ha sido retirado o modificándolas cuando sea necesario. Esto se debe a que la WWW es un entorno dinámico, que cambia constantemente según las necesidades de sus propietarios y usuarios (Calaresu y Shiri, 2015).

Si bien las herramientas actuales pueden transportar al usuario a diferentes sitios web, crear información, gestionarla o transformarla, no tienen como entender lo que esta significa y cómo se vincula a otros posibles resultados o datos. La causa radica en la ausencia de metadatos semánticos y ontológicos que proporcionen información adicional para comprender contenidos, significados y la relación entre los datos. Lo que a su vez facilitaría la comunicación y el intercambio de información entre diferentes entidades o sistemas, por ello tienen un comportamiento autónomo completamente reducido (Castells, 2005).

En este sentido, un esquema de operación web podría estar compuesto por agentes software dotados de cualidades tales como autonomía para interactuar con el entorno y tomar decisiones; adaptabilidad para aprender las preferencias del usuario, de otros agentes y de los orígenes de la información, y comunicación para inferir las necesidades y objetivos que tiene el usuario (Peis et al., 2003). Así tendría la capacidad de navegar y ejecutar acciones por sí mismo, con el objetivo de ahorrarle tiempo al usuario y mejorar los resultados. Para ello, es necesario que cada recurso web cuente con una descripción que brinde representación procesable y entendible por programas que puedan asistir, e incluso reemplazar, a los usuarios en tareas rutinarias o de difícil ejecución (Lago y Cacheiro, 2008)

Este es el sentido de Semantic Web, un proyecto que pretende que las páginas en internet contengan algo más que datos para que interpretan los computadores. Es el caso de una página codificada a través de HTML, cuyas *etiquetas* le indica al computador cómo mostrar el conjunto de caracteres en pantalla. Debido a esto, el creador de internet, Tim Berners-Lee, le apuesta a que los computadores puedan interpretar la información que contiene la web y a que sean capaces de tomar decisiones antes de mostrársela al usuario. De acuerdo con este visionario, lo que se espera es que en el futuro un usuario de Semantic Web pueda indicarle a su computador lo que necesita y que éste, al conocer sus preferencias, se encargara de llevar a cabo la tarea (Codina,

2003). La World Wide Web está creciendo rápidamente y a los usuarios les resulta cada vez más difícil obtener información útil y de manera eficiente. Debido a esto, esta iniciativa puede ayudar a reemplazar los denominados *enlaces web* por una *red de significados*, lo que hará que las búsquedas proporcionen contenido útil y de forma más eficiente (Fensel et al., 2003).

Para los usuarios, esta web representa que la información se puede acceder de manera más eficaz, porque actualmente se navega en un nivel bajo. Esto quiere decir que solamente es necesario señalar los enlaces y realizar búsquedas mediante palabras claves; pero, generalmente, estas acciones ofrecen información irrelevante. Un tipo de búsqueda diferente podría encontrar información que utiliza las palabras en contextos diferentes al que se pretende o que descubra enlaces que la utilicen y el abanico de respuestas será mayor. El manejo de ontologías ayuda a resolver este tipo de problemas, porque se centran en el contenido real de una página y de todas las relaciones que pueda tener en la red (Fensel et al., 2003). El modelo Semántico busca crear una web enlazada, que permita localizar, compartir o integrar información con mayor facilidad; incorporando metadatos semánticos y ontológicos que describan el contenido, el significado y las relaciones entre los datos. De esta manera es posible que los agentes puedan evaluar automáticamente la información y entregar resultados más precisos, mejorando el comportamiento autónomo que tienen hasta ahora (Castells, 2005). Los componentes más representativos de este modelo son los Metalenguajes y los estándares de representación XML, RDF, OWL, XML Schema y RDF Schema.

Para lograrlo, se requiere una infraestructura compuesta por un nuevo lenguaje de codificación de páginas, un nuevo lenguaje de marcado, gestión metadatos¹ y nuevos programas y métodos software para representar el conocimiento, los cuales requerirán de ontologías como procedimientos normalizados para representarlo (Codina, 2003). Además, es necesario que los propietarios de la información, las páginas y demás recursos alojados en la web, utilicen metadatos con el fin de que los motores de búsqueda y el software especializado los puedan encontrar para tomar decisiones.

¹ Muchas páginas actuales no tienen semántica, sintaxis, ni normas que agrupe las diversas plataformas de metadatos. Para ello se utiliza la norma RDF.

Algunos autores ya han propuesto modelos de arquitecturas por capas para Semantic Web, tal como se observa en la Figura 2.

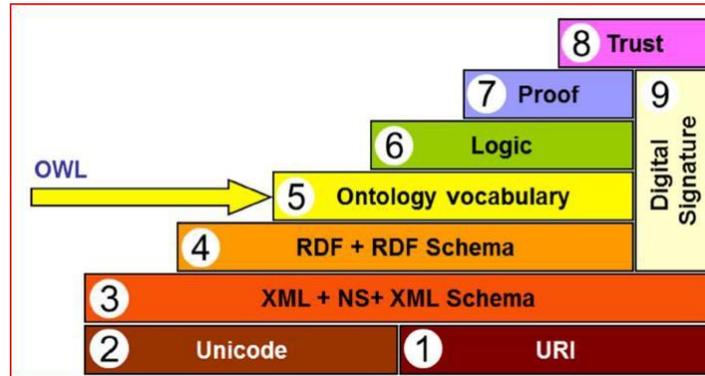


Figura 2. Capas de la arquitectura de Semantic Web (Berners, Hendler y Lassila, 2001).

- Universal Resource Identifier URI. Una cadena formateada que sirve como medio para identificar la ubicación de recursos físicos o abstractos.
- Unicode. Provee un único número para los caracteres, independientemente de la plataforma, programa o lenguaje subyacente. Es requerido para estándares tales como XML, Java, LDAP, WML, etc.
- XML + NS + XMLS. El lenguaje XML permite separar el contenido de la presentación de los documentos a través de la estandarización de formatos de datos, y proporciona la sintaxis superficial para los documentos estructurados. XMLS es un lenguaje que permite definir la estructura de los documentos XML.
- RDF + RDFS. RDF provee un modelo para describir recursos en la web, que son identificables únicamente a través de una URI y que describen la información de manera que pueda ser entendida por las aplicaciones de los computadores. RDFS es el vocabulario que permite describir las propiedades y las clases de los recursos RDF; adicionalmente, aporta la semántica para establecer jerarquías de generalización entre las propiedades y clases. Actualmente, RDF es el modelo más extendido en la comunidad de Semantic Web. Su elemento básico de construcción es la *Sentencia*, que está constituida por dos nodos (*Sujeto* y *Objeto*), que se unen por medio de un arco llamado *Predicado*. En este modelo, los Nodos representan recursos y los Arcos propiedades (Castells, 2005).
- Ontology Vocabulary. OWL es un lenguaje para construir ontologías a través de la descripción detallada de propiedades y clases, tales como Relaciones entre clases,

cardinalidad, igualdad, tipologías de propiedades más complejas y caracterización de propiedades o clases enumeradas.

- Logic, Proof y Trust. Son capas aun no exploradas significativamente.
- Digital Signature. Las capas 6, 7 y 8 utilizan firmas digitales como componentes de seguridad.

Por su parte, Cardoso y Pinto (2015) describen la dirección de las investigaciones futuras con respecto a Semantic Web (Web 3.0), como un pilar fundamental para el desarrollo de nuevos Sistemas de Información. Afirman que, para lograr, se requiere un lenguaje que permita la descripción de conceptos de un dominio dado y la creación de ontologías, como lo hace OWL. Esto conlleva a que en la web se puedan compartir documentos y datos, y a que la búsqueda de información sea más fácil, reutilizable y confiable. OWL es el lenguaje que pueden utilizar las aplicaciones para mejorar los motores de búsqueda y la gestión del conocimiento.

Codina (2003) afirma que las posibilidades de implementar Semantic Web son reducidas a corto y mediano plazo, pero que actualmente se incrementan las posibilidades debido a que se han desarrollado numerosas aplicaciones sobre estas conceptualizaciones. Entre sus argumentos, y en relación con las ontologías, describe que si la Inteligencia Artificial tiene varias décadas de fracasos, por lo menos en su hipótesis fuerte de lograr que los computadores se acerquen a algo parecido a la función humana de pensar, ¿por qué van a tener éxito ahora así de repente? Todavía no está convencido de que estos desarrollos serán exitosos, pero se muestra esperanzado, siempre y cuando haya un cambio de paradigma en las Ciencias computacionales y en las teorías de la documentación. Aún que sobre esta última, argumenta que los documentalistas desde hace décadas podrían estar construyendo una Semantic Web. La cuestión es que si esto es cierto, ¿por qué no han generado aún el cambio de paradigma que él solicita?

En su libro, Codina, Marcos y Pedraza (2009) describen los tipos de cambios que podría aportar Semantic Web: 1) en relación con la Inteligencia Artificial, donde esperan que las tareas que realizan los humanos exitosamente a través del razonamientos, la hagan los computadores a través de sistemas de metadatos, ontologías y lógica formal; 2) en relación con el procesamiento de la información, es decir, el impacto que puede tener

con los Sistemas de Información documental. Estos autores no ven un futuro muy promisorio en el primer aporte, pero sí muchas posibilidades en el segundo.

Rodríguez, Merlino y García (2015) presentan la evolución de los métodos de extracción de conocimiento de relaciones semánticas para la web. Éstos son técnicas mediante las que los procesos automatizables analizan fuentes de información no-estructurada (como textos escritos en lenguaje natural) y extraen el conocimiento embebido, para representarlo de manera estructurada y manipulable en procesos de razonamiento automático. Los autores presentan un análisis comparativo de la evolución de los métodos de extracción del conocimiento creados para la web (ver Figura 3). Su objetivo es demostrar cómo, entre 2005 y 2015, cada nuevo método mejora las técnicas de extracción de sus antecesores. De acuerdo con este aporte, se observa el desarrollo del trabajo en Semantic Web. Esto sustenta lo que previamente argumentaba Codina (2003), en el sentido de que a corto y mediano plazo no se tendrán resultados relevantes. Pero también se observa que el cambio de paradigma que él solicita sí se está dando poco a poco.

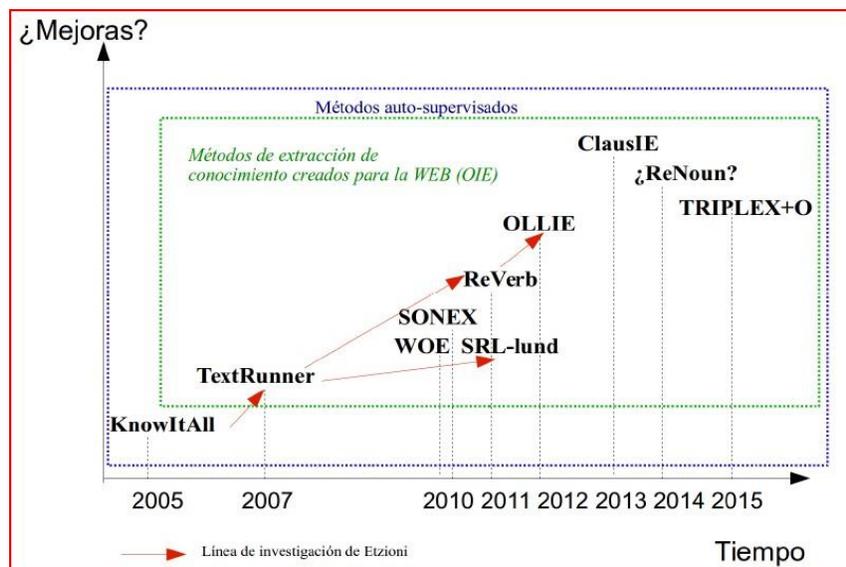


Figura 3. Desarrollo de los métodos de extracción de conocimiento para la web (Rodríguez, Merlino y García, 2015)

Desde las áreas de la educación también se están desarrollando propuestas para Semantic Web, entre las que vale la pena mencionar el trabajo de Villarreal et al. (2015), quienes proponen un modelo de Repositorio Semántico de Objetos de Aprendizaje (RSOA) basado en ontologías para entornos virtuales. Su idea es poder compartir el

	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

conocimiento y los contenidos multimedia e instruccionales de los cursos *e-learning*, que tienen la característica de ser reutilizados y referenciados en diferentes contextos. Para el repositorio se haría uso de la arquitectura SOA, dado que su enfoque hace que los contenidos sean accedidos e interpretados por agentes, no solamente por personas, llevando esto a un nivel de interoperabilidad semántica.

Para Mora et al. (2015) Semantic Web es una herramienta que se puede aprovechar en los procesos docencia-aprendizaje para estructurar, almacenar y compartir los contenidos sobre los diferentes cuerpos de conocimiento que se procesan en las instituciones de educación. Para ellos, una de las líneas de investigación docente más activas es desarrollar nuevas formas de gestión de la información en internet. El reto consiste en crear inteligencia colectiva a partir de interacciones web, en lugar de considerar las interacciones de los estudiantes de forma individual. Estas herramientas permiten que los profesores transmitan de mejor manera los conocimientos y que los estudiantes los asimilen más rápido. Así que el cuerpo del conocimiento puede ser explicado desde diferentes puntos de vista obteniendo el mismo resultado.

Otro referente con relación al desarrollo y uso de Semantic Web es el que publican Pandey y Panda (2015) acerca de las diferentes soluciones (herramientas, servicios y proyectos) semánticas desarrolladas, que pueden utilizarse en diferentes áreas del conocimiento, y las soluciones semánticas para bibliotecas digitales basadas en tecnologías web semánticas. Entre las herramientas usadas para la construcción de esta bibliotecas, los autores mencionan: almacenamiento para RDF (Sesame, Jena, YARS), razonadores (KAON, Racer), editores (Protege, SWOOP, MarcOnt Portal) y wikis Semánticos (Makna, SemperWiki). Asimismo, mencionan los desafíos que tiene que afrontar Semantic Web:

- Enfocarse en los motores de búsqueda, ya que hay muy pocos desarrollados de dominio público con enfoques de navegación y búsqueda semántica.
- Dado que la mayoría de las páginas web están desarrolladas en HTML, se requieren estándares semánticos para uso general, que se apliquen a las páginas y así los computadores puedan entender el significado asociado a los conceptos de búsqueda.
- Existen pocos estándares globales y *frameworks* probados para mapeo de dominios y representación del conocimiento, y pocos metadatos en los recursos web.

- Superar la debilidad en la disponibilidad de ontologías específicas de dominio formal, porque son pocas las herramientas desarrolladas para ello. Pero destacan SAPHIRE, un sistema de información en salud basado en semánticas y capaz de rastrear y evaluar situaciones y ocurrencias que podrían afectar la salud pública.

Un valor relevante de las bibliotecas digitales en comparación con la web tradicional, es que en ellas se puede implementar más fácilmente las tecnologías Semantic Web, debido a que las bibliotecas tienen metadatos para catalogar sus recursos.

En el área de la Ingeniería del Software, Aguilar y Portilla (2015) proponen un *framework* basado en ODA para describir y componer servicios web semánticos (FODAS-WS), el cual es un marco de referencia que sirve para diseñar y describir servicios Web semánticos automáticamente. El caso de estudio que presentan consiste en el diseño y la implementación de un servicio web semántico, que básicamente es un sistema recomendador, es decir, que ayuda a emparejar a usuarios con productos. Según los autores, este sistema está en capacidad de recomendar dos tipos de productos: objetos de aprendizaje y documentos web, cuyos metadatos se encuentran almacenados en repositorios semánticos de objetos de aprendizaje y en repositorios semánticos de archivos web.

Actualmente, la web es principalmente un sistema donde se almacenan y recuperan datos, sin embargo, el rápido desarrollo, la Sociedad de Conocimiento y la globalización, están exigiendo que se convierta en un sistema para *hacer cosas* y no solamente para *encontrar cosas*. Esto será posible a través de la construcción de un lenguaje semántico, en el que se puedan utilizar reglas acordadas, significados, normas y ontologías. Por lo tanto, los programadores podrán crear sistemas capaces de transmitir un significado, de manera que las maquinas puedan entender lo que se busca. Por ello, Semantic Web es alternativa revolucionaria para los actuales métodos de búsqueda. Aunque es posible que haya algún tipo de resistencia a esta estandarización, es necesaria debido a la cantidad de dispositivos, aplicaciones y lenguajes existentes, y al volumen de datos que existen actualmente. Con esta tecnología se lograrán afianzar los sistemas semánticos, y no hacerlo es arriesgarse a que el Sistemas de Información más potente jamás concebido se ahogue en su propia popularidad (Fensel, 2003).

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

Deep Web

Cuando se investiga el concepto de la Deep WEB, es posible encontrar dos puntos de vista relacionados con su significado: 1) como aquella parte de internet que no es accesible desde los motores de búsqueda tradicionales y que, para lograrlo, se requiere herramientas especializadas. El centro de discusión de este enfoque es su funcionamiento, las herramientas existentes y el potencial como repositorio de datos para que los usuarios resuelvan sus necesidades. 2) como un medio inseguro, donde se refiere como el medio para alojar y desarrollar actividades que van en contra de la seguridad del mundo.

Pero, ¿qué es y cómo se desarrolla? Deep Web está compuesta por un sin número de bases de datos a las que se puede acceder a través de interfaces de consulta diferentes; la información que se puede obtener de ellas a modo de páginas web se genera de forma dinámica, en respuesta ante una búsqueda; no existen páginas estáticas con vínculos a otros sitios en espera de ser invocados; y, por su estructura, y en algunos casos, es común que los buscadores tradicionales no puedan indexarlas y mantenerlas disponibles. Para lograrlo, sería necesario repensar el modo de ejecución de los mismos y sus *crawlers*, porque debe tener la capacidad de crear con antelación las consultas apropiadas, que arrojen páginas adecuadas y coherentes; dicha tarea requiere comprender el contenido de los sitios consultados para focalizar las consultas (Shestakov, 2008). Deep Web no solamente se debe a la existencia de registros alojados en bases de datos, sitios pornográficos, pedófilos, dictatoriales, expendedores de drogas, prostitución, sicarios y demás, porque aproximadamente el 90% de esta red se alimenta en gran medida de páginas tradicionales y legalmente construidas, aunque no registradas, ya sea por decisión de su propietario o simplemente por la estructura empleada en su construcción.

Cuándo un usuario realiza una consulta tradicional a través de un buscador de internet, lo que hace el motor de búsqueda es tomar las palabras digitadas y recuperar aquellas páginas que se hallan indexadas a la fecha en su base de datos. La búsqueda traerá como resultado un porcentaje determinado que concuerde con lo buscado. Pero, con seguridad, el motor recupera estas páginas de la web superficial, es decir, aquella a la que tiene acceso la gran mayoría de personas en el mundo. Para hacerse a una imagen

de esta situación, Google, aunque ha mejorado su funcionalidad utilizando un sistema de indexado y *crawling* basado en la popularidad de las páginas o documentos (aquellas más visitadas tendrán prioridad para salir en los primeros lugares del listado de resultados), solamente accede al 16% del total de páginas en la web superficial (NEC Laboratories America: <http://www.nec-labs.com/>). Esto quiere decir que los usuarios solamente encuentran el 0.03% de las páginas disponibles, porque el porcentaje restante se aloja en Deep Web (Bergman, 2001). Entonces, se puede decir que esta web es donde se alojan bases de datos y páginas no indexadas y no consultadas a través de motores de búsqueda convencionales.

Bergman (2001) concluye que los motores de búsqueda deben evolucionar para poder acceder tanto a la web superficial como a Deep Web, a la vez que deben satisfacer las consultas de los usuarios. El autor afirma que esto se podría lograr a través de tecnologías de consulta dirigida, donde se involucre la *mega* búsqueda de sitios Deep Web y la *meta* búsqueda de la web superficial. Porque actualmente, aunque existen medios especializados para recolectar información específica, no son útiles para encontrar toda la información disponible sobre un tema o área. Según Bergman, esto sugiere una división del mercado de búsqueda de información en internet, es decir, directorios de búsqueda, que ofrezcan información seleccionada y específica desde la web superficial para satisfacer necesidades de búsqueda populares; motores de búsqueda, para consultas de nivel superior en esta web; e *infohubs*, para la información Deep Web que proporcione respuestas donde son imperativas la comprensión y la calidad.

En cierta forma, esto corresponde a la tecnología propuesta por Brightplanet (<http://www.brightplanet.com/>) y desarrollada hace tiempo para socavar en Deep Web y traer los resultados de vuelta a la web superficial. Es decir, que al identificar donde residen las bases de datos apropiadas, se puede realizar una consulta simultánea en cada una de las fuentes para recolectar únicamente los resultados deseados con una alta precisión. Esta tecnología utiliza múltiples hilos de múltiples fuentes de consulta y documenta las descargas, luego indexa todos los documentos recuperados y, mediante algoritmos específicos, los clasifica por relevancia de uso (Bergman, 2001).

Aplicando esta tecnología, BrightPlanet ha aprovechado Deep Web y las debilidades de los motores de búsqueda tradicionales, para encontrar los resultados esperados por los usuarios. Debido a que en su proceso de indexación de páginas web, los motores de búsqueda solamente almacenan las palabras mencionadas con más frecuencia, a la vez que su localización y unos pocos metadatos, esta empresa desarrolla herramientas especializadas para clientes que requieren consultar información sobre temas específicos. Al tener todas las palabras contenidas de una página web en un amplio conjunto de sitios, utiliza tecnologías con capacidades analíticas, ayudándoles a sus clientes a visualizar, analizar y crear inteligencia de grandes conjuntos de datos (Pederson, 2013).

Este tipo de empresas ha logrado almacenar con los años grandes repositorios de datos sobre diferentes sitios web, que utilizan para entregar información a sus clientes. En tal sentido, son los clientes quienes definen su objetivo de búsqueda, y es a través de estas herramientas, desarrolladas con tecnología de web superficial, que logran obtener los resultados esperados desde Deep Web. Un ejemplo de esta relación es lo que ha estructurado la plataforma OpenPlanet. Allí se separa la exploración y los componentes analíticos de la recopilación de datos y el análisis. En otras palabras, no importa de donde provengan los datos, el objetivo es tomar todos los que defina el cliente e iniciar un proceso para generar los resultados esperados por el mismo. En la Figura 4 se observa el proceso de una exploración desde diferentes fuentes: web superficial, Deep Web, archivos de fuentes RSS, blogs, tweets, correos electrónicos, etc.

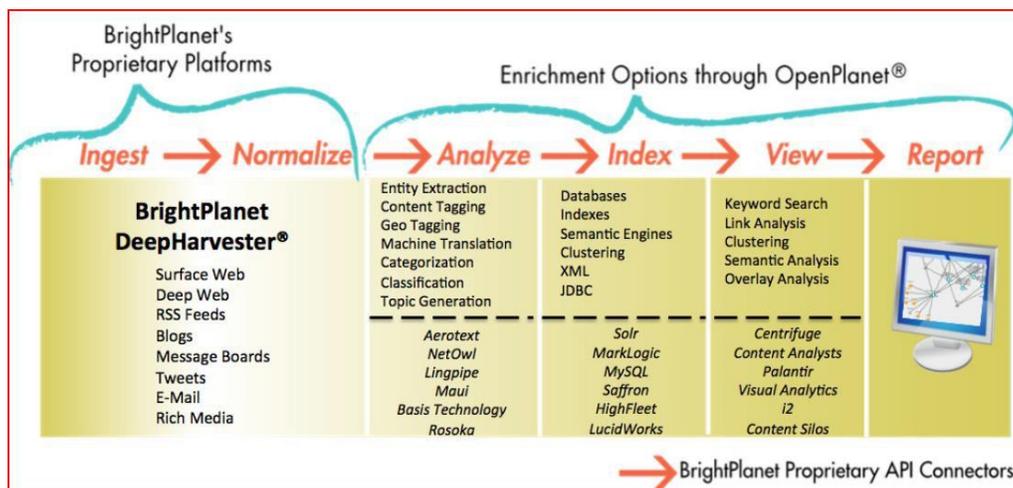


Figura 4. Flujo de trabajo de la plataforma OpenPlanet (Pedersón, 2013)

De forma paralela a estas descripciones y aportes de diversos autores, también es necesario hablar de los aspectos positivos y negativos que se publican alrededor de Deep Web. Aquí es importante destacar que en internet se encuentran muchas y diversas fuentes en las que se hace referencia Deep Web como una web oscura o invisible, pero generalmente son interpretaciones malintencionadas (Pederson, 2013). Claro está, no se puede negar la existencia de peligros que pueden afectar la seguridad de la información y los equipos de los usuarios que navegan por allí, pero lo mismo podría decirse de cuando lo hacen en la web superficial.

En el documental de Mike Radford, *Horizon: Inside the dark Web*, este director describe los pros y los contras de navegar en la web y la forma en que las personas son vigiladas, registradas y analizadas. Desde el surgimiento de internet ha existido una preocupación latente por la privacidad de la información, ya sea personal, gubernamental o empresarial, y que ha sido y es usada por los estados para saber quiénes son los usuarios, qué hacen y si son un riesgo para la sociedad y el mundo. Por otro, también existen personas y organizaciones que utilizan esa información para venderla a otras empresas que necesitan conocer su estilo de vida, qué les gusta y lo que posiblemente podrían comprar. Marc Victoria (2015) hace referencia a una herramienta desarrollada por Financial Times, que puede calcular cuánto podrían pagar ciertos grupos interesados en la información que circula en internet.

Por su parte, los gobiernos le han apostado al tema de salvaguardar la privacidad de su información, con el fin de defenderse de posibles ataques terroristas u otras actividades que afecten a la sociedad. La cuestión que es necesario discutir es qué tanta privacidad se puede perder al navegar en la web. Deep Web se diseñó con el objeto de mantener el anonimato, lo que puede verse como positivo para mantener la privacidad y tener libertad al momento de adentrarse en internet. Pero, lamentablemente, este tipo de estructuras es visto y aprovechado por otros para desarrollar actividades ilícitas, y es desde este punto de vista supone una amenaza para los estados.

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

5. CONCLUSIONES

La mayoría de autores consultados afirman que Semantic Web es un paso evolutivo de la red actual y que su masificación generará avances en relación con los modos de acceso y representación del conocimiento. Pero, para lograr un impacto real se necesita un cambio en la cultura del diseño y desarrollo de páginas web, porque actualmente los sitios alojados en internet carecen de la información semántica requerida. Las páginas deben contener metadatos que les permitan a las herramientas analíticas actuar de manera inteligente y obtener los resultados que esperan los usuarios al realizar las consultas. Por otro lado, algunos de estos autores se muestran escépticos con respecto a un avance rápido de Semantic Web, sobre todo los relacionados con la documentación. Ellos la comparan con el desarrollo de la Inteligencia Artificial y afirman que si después de tantos años de desarrollo no ha podido lograr el avance esperado, esta web tendría dificultades semejantes. Esto se podría evitar con un cambio de paradigma en relación con la estructuración semántica que requiere internet.

Otro aspecto a resaltar es que el progreso de Semantic Web es lento debido a que la mayoría de esfuerzos son independientes. Por ello se requiere de un cambio en la legislación, donde se estandarice el uso de la semántica en todos los sitios web. El desarrollo se visualiza en intentos por estructurar sitios aislados, pero no tanto en motores de búsqueda.

Por otro lado, además que se debe buscar el mejoramiento en el desarrollo de Semantic Web, es claro que Deep Web tiene un alto potencial debido a que en ella se aloja más de las tres cuartas partes de los recursos de información existentes en internet. Pero el caso de ambas es similar, en el sentido de que los esfuerzos por desarrollarlas son mayormente independientes, y se requiere un trabajo más de comunidad.

Aunque los motores de búsqueda tradicionales han progresado en el desarrollo de funcionalidades, todavía no son la herramienta que requieren los usuarios en Deep Web, por lo que se necesita mayor trabajo para alcanzar este territorio y brindar

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

resultados más adecuados en cada consulta de los usuarios. La idea es que no tengan que depender de herramientas especializadas para acceder el volumen de información que está depositada en ella.

Actualmente, las probabilidades de encontrar información de calidad en internet son tres veces más probables en Deep Web que en la web superficial, por lo que el reto para la comunidad es buscar formas de masificar el acceso a ella.

Desde el punto de vista de la seguridad, Deep Web tiene una arquitectura en la que se puede navegar de forma anónima. Esto es útil para salvaguardar la privacidad y transferencia de información de los usuarios, debido a que tiene plataformas y/o redes creadas para tal fin, evitando que se puedan vulnerar sus derechos. La otra cara de la moneda en relación con esta privacidad, es que se presta para realizar actividades ilícitas, lo que dificulta el trabajo de proteger la información de los usuarios.

Como se observa en este artículo, tanto Semantic Web como Deep Web tienen una realidad diferente. La primera es un desarrollo que intenta aprovechar la información a la que los buscadores no pueden llegar, de tal forma que sean los sitios y los computadores los que *dialoguen* para responder las consultas de los usuarios. De esta forma busca entregar los mejores resultados en cada búsqueda que realicen. Mientras que la segunda, con todo el volumen de información que posee y con todo el potencial de aprovechamiento, se encuentra inmersa en mitos y leyendas que no han posibilitado su masificación. Actualmente está restringida a comunidades seleccionadas, debido en parte a que los buscadores tradicionales no la acceden y a que el anonimato en la navegación no es visto desde el punto de vista de la seguridad de los estados y organizaciones.

Se recomienda realizar estudios prácticos sobre la plataforma “Tor” que permitan evidenciar con mayor grado de detalle, la funcionalidad y disponibilidad de los diferentes tipos de información almacenados en ella; para ello se hace necesario profundizar el conocimiento en la estructura URL.onion que es completamente diferente a la que conocemos o manejamos cotidianamente, o en su defecto, explorar portales que suministren información de acceso a las diferentes páginas.

REFERENCIAS

- Aguiar, C. y da Costa, K. (2013). O que a Deep Web pode oferecer além da Surface Web. Proceedings XV Semana de Informática e XII Mostra de Trabalhos de Iniciação Científica (pp. 1-5). Paranavaí, Brazil.
- Aguilar, J. y Portilla, O. (2015). Framework basado en ODA para la descripción y composición de servicios web semánticos (FODAS-WS). Latin American Journal of Computing 2(2), pp. 15-24.
- Bergman, M. (2001). [White Paper: The Deep Web: Surfacing hidden value](#). The journal of electronic publishing 7(1), online [Aug. 2015].
- Berners, T., Hendler, J. y Lassila, O. (2001). The semantic web. Scientific American, 279(5), pp. 34-43.
- Calaresu, M. y Shiri, A. (2015). Understanding Semantic Web: A conceptual model. Library Review 64(1/2), pp. 82-100.
- Cardoso, J. y Pinto, A. (2015). The Web Ontology Language (OWL) and its Applications. Encyclopedia of Information Science and Technology.
- Castells, P. (2005). La web semántica. En Bravo, C y Redondo, M. (Eds.), Sistemas interactivos y colaborativos en la web (pp. 195-212). España: Ediciones de la Universidad de Castilla-La Mancha.
- Chang, C. et al. (2006). A survey of Web information extraction systems. IEEE Transactions on Knowledge and Data Engineering 18(10), pp. 1411-1428.
- Ciancaglini, V. et al. (2014). [Deepweb and Cybercrime - It's not all about TOR](#). A Trend Micro research paper. Online [Aug. 2015].
- Clarke, I. (1999). A Distributed, decentralised information storage and retrieval system. Division of Informatics. Edinburgh University.
- Codina, L. (2003). La Web semántica: Una visión crítica. El profesional de la información 12(2), pp. 149-152.
- Codina, L., Marcos, M. y Pedraza, R (2009). Web Semántica y Sistemas de Información documental. España: Trea.
- Fensel, D. et al. (2003). Spinning the Semantic Web. USA: MIT Press.
- He, B. et al (2007). Accessing the deep Web: A survey. Communications of the ACM 50(2), pp. 94-101.
- Lago, B. y Cacheiro, M. (2008). Web Semántica Web 3.0. En Marchessou, F. (Ed.), La Web 2.0 y el mundo educativo. XIII congreso Internacional de Informática educativa. Madrid, España.
- Mora, H. et al. (2015). La Web Semántica como herramienta para el apoyo a la docencia. Memorias XIII Jornadas de Redes de Investigación en Docencia Universitaria (pp. 1-14). Alicante, España.
- NEC Laboratories America. <http://www.nec-labs.com/>
- O'Brien, D. (2005). UK digital rights group sets up. BBC News, 9 September.

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

- Pandey, S. y Panda, K. (2015). Semantic solutions for the digital libraries based on semantic Web technologies. *Annals of Library and Information Studies* 61(4), pp. 286-293.
- Parundekar, R., Knoblock, C. y Ambite, J. (2010). Linking the Deep Web to the Linked DataWeb
- Pederson, S. (2013). Understanding the Deep Web in 10 minutes. White paper: Birghtplanet.
- Peis, E. et al. (2003). Análisis de la web semántica: Estado actual y requisitos futuros. *El profesional de la información* 12(5), pp. 368-374.
- Radford, M. (Productor). (2014). *Horizon: Inside the dark Web*. BBC. London, UK.
- Rodríguez, J., Merlino, H. y García, R. (2015). Revisión sistemática comparativa de evolución de métodos de extracción de conocimiento para la Web. *Memorias XXI Congreso Argentino de Ciencias de la Computación* (pp. 1-10). Junín, Argentina.
- Sherman, C. & Pric, G. (2003). The Invisible Web: Uncóvering sóurces search engines can't see. *Library Trends* 52(2), pp. 282-298.
- Shestakov, D. (2008). *Search interfaces on the web: Querying and characterizing*. Tesis doctoral. Universidad de Turku, Finlandia.
- Staab, S. & Studer, R. (2008). *Handbook on Ontologies*. USA: Springer.
- Varde, A, et al. (2009). Knowledge discovery over the Deep Web, Semantic Web and XML. *Lecture Notes in Computer Science* 5463, pp. 784-788.
- Victoria, M. (2015). *Deep Info! Buscador de datos personales en la Deep Web*. Trabajo de fin de grado. Escola Superior Politècnica UPF, Barcelona.
- Villarreal, Y. et al. (2015). RSOA basado en ontologías para entornos virtuales en la educación superior panameña. *Virtual educa*. Guadalajara, México.
- Zaino, J. (2009). [Why the Deep Web needs the Semantic Web](#). Semanticweb.com. Online [Sept. 2015].
- Zillman, M. (2014). [Deep Web research and discovery resources 2015](#). LLRX.com. Online [Sept. 2015].

APENDICE (Artículo científico)

Semantic Web y Deep Web: Análisis de una realidad

Semantic Web and Deep Web: Analysis of a reality

Edgar Serna M., Diego Villada V., Andrés Jiménez V.

Resumen

Se puede afirmar que la información depositada en internet está distribuida en una especie de edificio y que para accederla se debe navegar por pisos llenos de ella. La cuestión es que los navegadores tradicionales no tienen la potencia para circular por todos esos pisos y se restringen sólo a unos cuantos. Debido a esto, los usuarios no encuentran respuestas que satisfagan completamente sus búsquedas. Los primeros pisos de este edificio, que podrían considerarse los más básicos, contienen un porcentaje muy pequeño de la información en la web, y son a los que recurren los buscadores, es decir, es una navegación superficial. El resto de pisos, la mayoría, están por fuera de su alcance y el contenido permanece oculto en cada consulta. Para lograr equiparar un poco esta realidad se han desarrollado dos movimientos: Semantic Web y Deep Web. Ambos tienen la intención de permitirles a los usuarios alcanzar la mayor parte de la información en la web. El primero desde un punto de vista semántico y el segundo ofreciendo privacidad al momento de navegar. En este artículo, se presenta un análisis a la realidad de ambas tecnologías y se desarrolla la conceptualización necesaria para mejorar la comprensión de ellas.

Palabras Clave: Ciberseguridad, gestión del conocimiento, Internet, motor de búsqueda, navegador.

Abstract

It can be said that information lodged on the Internet is divided into a kind of building and to access it must navigate floors full of it. The issue is that traditional browsers does not have the power to drive through all those floors and are restricted to only a few. Because of this, users do not find answers that fully meet their searches. The first floors of this building, which could be considered the most basic, contain a very small percentage of the information on the web, and are those who use the search engines, it means, it is a shallow navigation. The remaining floors, the most, are outside its scope and content remains hidden at every visit. To achieve equate a little this reality a little have been developed two movements: Semantic Web and Deep Web. Both are intended to enable users to reach the most part of the information on the web. The first from a semantic point of view and the second offering privacy when browsing. In this article, is presented an analysis of the reality of both technologies and is develop the conceptualization necessary to improve understanding of them.

Keywords: cybersecurity, knowledge management, Internet, web search engine, web browser.

INTRODUCCIÓN

Cuando se piensa en la web, la mayoría de personas hace referencia a la web visible, es decir, las páginas que se pueden acceder directamente a través de su enlace o desde los motores de búsqueda. Sin embargo, esta imagen de la web solamente representa el 4% de todo el contenido en internet. La web se ha convertido rápidamente en una entidad compleja y se estima que cada año se añade un *zetabyte*, o más, de contenido. Navegar en esta cantidad de información representa un desafío, especialmente porque a través de los motores de búsqueda tradicionales no se accede fácilmente a ella. En términos generales, a la web que se puede acceder de esta forma se le conoce como *Semantic Web*, y a la parte *oculta* o *invisible*, que se accede solamente a través de herramientas especializadas o páginas con control de acceso, se le conoce como *Deep Web*. Esta última recibe su nombre debido a la incapacidad de los navegadores

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

tradicionales para acceder a la información no-catalogada o indexada que se encuentra en ella. Esto se debe a que los motores de búsqueda todavía no incorporan los nuevos estándares de indexación y no tienen la capacidad de acceder a páginas dinámicas. Otras razones van desde el bloqueo impuesto desde la estructura de sus sitios, porque están protegidos por contraseña de seguridad mediante variables *no follow* para evitar el seguimiento de la huella digital, o simplemente porque contienen formatos que no pueden ser interpretados por los robots *crawlers*, encargados buscar e indexar la información. En estos casos, el motor de búsqueda no encuentra el código que leen los robots en el contenido asociado para realizar referencia estática, o simplemente no pueden acceder a las bases de datos protegidas, por lo que pasan de largo ante estos sitios (Shestakov, 2008).

Ian Clarke (1999) propuso un sistema de información de almacenamiento y recuperación distribuido y descentralizado, o como algunos podrían afirmar, una nueva y revolucionaria forma de utilizar Internet sin ser detectado. Es decir, con este software, cualquier persona podría conversar en línea, o leer o crear una página web, o compartir archivos con anonimato casi total. Su objetivo era evitar los estrictos controles que suponía la red en ese entonces, porque para él Internet podía ser controlado con mayor rapidez, de forma más completa y más barata que con los sistemas de comunicación de la época, y su software pionero tenía la intención de cambiar eso.

La red moderna a menudo es considerada como un milagro tecnológico, en el que aparentemente todo funciona sin contratiempos. Muchos usuarios piensan que cuando realizan una búsqueda a través de un motor cualquiera, van a tener acceso a toda la información en todas las páginas web. La realidad es que ese buscador solamente accede a una fracción muy pequeña de la verdadera profundidad de la web, y trae a la superficie pequeños fragmentos de la información disponible. Aunque no se tiene forma de estimar cuán grande y profunda es la web, sí se tienen certeza de que el área en la que navega la mayoría de internautas es una fracción muy pequeña de ella. Al resto se le ha denominado de diversas formas: *darkweb*, *deep web*, *surface web*, *invisible web*, *dark address space*, *murky address space*, *dirty address space*, y esas metáforas sólo hacen que internet se convierta en más insondable y misteriosa cada vez. La razón es que no todos estos eufemismos significan lo mismo: mientras que *darknet* es una red en línea, como Freenet, que se oculta a los no-usuarios, con todo el potencial que implica en cuanto comportamiento transgresor, gran parte de *deep web*, atemorizante como suena, consiste del consumo e investigación de datos sin complicaciones, pero que está más allá del alcance de los motores de búsqueda tradicionales. Por otro lado, *dark address space* se refiere a menudo a las direcciones de Internet que, por razones puramente técnicas, simplemente han dejado de funcionar.

Pero, en cierto sentido, todos forman parte de una misma imagen, en la que más allá de los confines de la vida en línea de la mayoría de personas, existe por ahí otra inmensa internet, utilizada por millones, en gran parte ignorada por los medios pero bien entendida solamente por unos pocos científicos computacionales. A finales de los años 90, Bergman (2001) realizó una investigación para tratar de medir el tamaño de internet. Partió de la hipótesis de que, conociendo la web de acceso cotidiano, era probable que la otra web fuera dos o tres veces más grande. Pero, luego de más de diez años de trabajo, la inmensidad de la misma todavía lo deja sin aliento, y sigue girando sobre supuestos y sin descubrir lo que se propuso entonces. Solamente se ha acercado con algunas cifras y suposiciones, como que actualmente esta web es 400 a 550 veces más grande que la Semantic Web, que es la de más rápido crecimiento en nueva información, que el valor de su contenido es inconmensurable y que los buscadores tradicionales solamente tienen acceso al 0,03% del total de las páginas disponibles en internet.

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

Pero su tamaño no es el único problema, también hay que enfrentarse a los sitios bloqueados. Muchas veces la motivación es comercial, porque han invertido mucho tiempo y dinero como para dejar la información libre, otras veces es por intereses delictivos, cuando el contenido del sitio está dirigido a cometer actos ilícitos y no le conviene salir a la luz pública. De hecho, internet siempre ha estado impulsado por el deseo de mantener el secreto como una idea de transparencia. La red es una creación conjunta del Departamento de Defensa de los Estados Unidos y WELL, una comunidad de contracultura americana y una de los primeras y más influyentes en la web. Ambos grupos tenían razones para construir un entorno oculto o semi-oculto en línea, lo mismo que otro abierto. Por otro lado, todavía existen rincones secretos de internet donde perdura esta extraña alianza entre liberales y militares. Pero, para la mayoría de personas y empresas, la web es un universo en el que pueden comprar, vender, intercambiar y conocer sitios y personas en todo el mundo. Por todo esto, parece probable que Internet permanecerá un tiempo más en su fase de fiebre del oro, por lo que en las grietas y esquinas de sus estructuras, seguirán apareciendo sitios oscuros y otros entornos en línea privados.

Estos y otros factores son los que generan mitos y creencias alrededor de ambas plataformas, y los usuarios han creado su propia imagen de ellas. Ya sea porque han tenido experiencias poco agradables en la navegación, o porque los comentarios son malintencionados, o porque se exagera el escenario que han vivido, en la red se encuentra todo tipo de referencias a la Semantic Web y a la Deep Web. Debido a que alrededor de ambos entornos de navegación se ha tejido un universo de mitos y creencias, es necesario aclararlos de la mejor forma.

El objetivo de este artículo es presentar un análisis a la realidad Semantic Web vs Deep Web, desde un punto de vista objetivo e imparcial. La idea es ofrecer pautas para que las personas comprendan las ventajas y desventajas de cada una, y presentarles los criterios de seguridad y de riesgo a los que se exponen cada día en la red. Este trabajo se encuentra estructurado de la siguiente forma: en la primera parte se detallan los trabajos relacionados con lo que se presenta, en la segunda se hace el análisis teórico a la realidad de Semantic Web y Deep Web y en la tercera se presentan las conclusiones respectivas.

TRABAJOS RELACIONADOS

Con el objetivo de alcanzar los mismos propósitos que se buscan en este artículo, varios autores han publicado sus trabajos en la literatura relacionada.

Varde et al. (2009) presentan un tutorial en el que describen un análisis al Web Mining, es decir, al conocimiento que se descubre en World Wide Web, especialmente con referencia a los últimos avances en la tecnología Web. Tratan los temas de Deep Web, Semantic Web y eXtensible Markup Language (XML), un medio de comunicación generalizado para la Web. También explican cómo cada uno de estos desarrollos apoya el descubrimiento de conocimiento a partir de los datos almacenados en Internet. Pero no analizan los mitos y realidades de cada uno de ellos, dejando la idea de que son armónicos y que conviven sin contratiempos en el ciberespacio.

Zaino (2009) afirma que Deep Web y Semantic Web están muy relacionados entre sí y se necesitan mutuamente. Para hacerse a una idea de si están compitiendo o son complementarias se llevó a cabo una investigación con varios autores representativos en este tema. La conclusión es que se complementan entre sí y que se necesitan para avanzar en sus metas diarias. Este autor analiza el mito en el que se afirma que ambas web están en continua disputa, y la realidad que demuestra es que no es cierto. Más allá de esto, no profundiza en los otros que se han difundido, lo que queda como investigación a continuar.

	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

De acuerdo con Parundekar, Knoblock y Ambite (2010), a pesar que Semantic Web gana terreno, grandes cantidades de información solamente están disponibles en las páginas web tradicionales. Los datos de fuentes como Surface Web y Deep Web necesitan ser publicados de forma estructurada en Semantic Web. Su enfoque no sólo relaciona las fuentes conocidas de Deep Web y Semantic Web, sino que también analiza los enlaces desconocidas para ambos dominios. Su idea es presentar un reporte de la cantidad de datos disponibles en internet. Pero, al igual que otros autores, este análisis no desmiente ni apoya la realidad de ambas superficies.

El artículo de Aguiar y da Costa (2013) tiene como objetivo presentar el lado oculto de Deep Web, conocida sólo por usuarios con amplia experiencia en navegar por la Web. De acuerdo con ellos, puede ofrecer una gran y enriquecedora variedad de contenidos de interés que apenas se encuentra en la web que se utiliza en el día a día. Además, es un área donde el anonimato es el principal objetivo. Este trabajo solamente analiza Deep Web, pero no clarifica si todo lo que se dice acerca de ella es cierto o no.

Para Ciancaglini et al. (2014), el término Deep Web se utiliza para designar una clase de contenido en Internet que, por diferentes razones técnicas, no está indexado por los motores de búsqueda tradicionales. Esta Web se refiere a una clase de redes que tienen como objetivo garantizar el acceso anónimo e indetectable al contenido y el anonimato para un sitio. Estos autores se centran en analizar la ciberdelincuencia que transita en estos sitios y la comparan con lo podría pasar en la WWW. Por último, presentan algunas técnicas que pueden utilizar los investigadores para supervisar proactivamente estas partes ocultas de internet. Su trabajo se orienta a analizar particularmente la inseguridad en la Deep Web, tratándola como una desventaja de esta red, pero no hace cruces con la que se presente en la Semantic Web.

En los últimos años, algunos de los motores de búsqueda más reconocidos han comenzado a escribir algoritmos para adentrarse en las partes más profundas de la web, tratando de encontrar archivos que no se encuentran del modo tradicional (Zillman, 2014). Estos procesos utilizan técnicas y algoritmos recientes que les permiten a los investigadores obtener una amplia cantidad de información que en Semantic Web no estaba disponible o es inaccesible. Esta investigación también demuestra que incluso los archivos más ocultos se pueden obtener a partir de búsquedas estructuradas. Aunque este autor describe una reseña del desarrollo de estas búsquedas, no se detiene a analizar la realidad de ambas plataformas.

ANÁLISIS A LA REALIDAD DE LA WEB

World Wide Web es hoy por hoy el repositorio más grande y completo que haya existido en la historia de la información. De acuerdo con la Unión Internacional de Telecomunicaciones, a finales de 2014 se estimaba que internet tenía 3.000 millones de usuarios, una cantidad que crece cada día y para la que es necesario ofrecer servicios adaptando las tecnologías actuales y mejorando los tiempos y la calidad de las respuestas. Para lograrlo, es necesario intervenir los modelos de desarrollo y mecanismos de escaneo e indexación que emplean los navegadores, para ofrecer con éxito las respuestas a las consultas de los usuarios. De lo contrario, la información contenida en los miles de sitios alojados en todo el mundo se podría desperdiciar (Calaresu y Shiri, 2015). El problema aquí es que el volumen de información en la Web crece cada día, pero las herramientas para consultarla no se desarrollan al mismo ritmo. Esto crea una especie de brecha entre la disponible y la que se puede acceder. Para la mayoría de usuarios esto no es representativo, porque sus necesidades de búsqueda se satisfacen fácilmente. Pero a los científicos y gobiernos se les dificulta realizar la mayoría de sus consultas de gestión del conocimiento, porque no es fácil acceder a la información con las herramientas actuales (Fensel et al., 2003).

Parte de toda la información documentada del planeta está disponible a través de los motores de búsqueda convencionales, que muchos denominan Web superficial. Sin embargo, existe un segmento mucho mayor que no se encuentra indexado por ellos y no es de fácil acceso, a esa Web se conoce ampliamente como Web profunda. Se estima que esta última contiene más de 2.5 billones de documentos, y que crecen a un ritmo de 7.5 millones por día. Además, debido a que la información en la red contiene enlaces de relación, los documentos que se encuentran dentro de otros documentos son más propensos a ser indexados en otros nuevos, y los buscadores tradicionales no los logran identificar. Para superar estos problemas, se ha desarrollado una nueva generación de motores de búsqueda en la cual se reemplaza el *link following* al azar, enfocándose en el rastreo y la indexación dirigida al tráfico de las páginas. Este enfoque genera mejores resultados en consultas simples, aunque aumenta la tendencia a pasar por alto información poco referenciada (Bergman, 2001).

Para comprender el proceso de ubicación de la información en internet, Marc Victoria (2015) detalla los niveles de clasificación de los sitios en la red, tal como se observa en la Figura 1. Al nivel 0 tienen acceso todos los usuarios de la web a través de motores de búsqueda convencionales, y las consultas muestran las páginas que fueron indexadas por el buscador. Este nivel cubre la porción más pequeña de la información documentada en internet. En los niveles 1 y 2 se puede consultar algunos sitios desde la Web superficial, y a partir del nivel 3 se ingresa al mundo de Deep Web, que es accesible a través de tecnologías o sitios especializados y en muchos casos solamente a través del pago del servicio.



Figura 1. Niveles de clasificación de las páginas web (Victoria, 2015)

Semantic Web

Cuando se realiza una búsqueda en la web, los navegadores realizan esta actividad en bases de datos especializadas que constantemente son actualizadas por *web crawlers*, una especie de robot que ejecuta consultas constantes entre enlaces de hipertexto para añadir la información que encuentran a sus bases de datos. Con esta operación proactiva agilizan la entrega de resultados y mantienen actualizadas las diferentes entradas, ya sea eliminándolas cuando el sitio ha sido retirado o modificándolas cuando sea necesario. Esto se debe a que la WWW es un

	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

entorno dinámico, que cambia constantemente según las necesidades de sus propietarios y usuarios (Calaresu y Shiri, 2015).

Si bien las herramientas actuales pueden transportar al usuario a diferentes sitios web, crear información, gestionarla o transformarla, no tienen como entender lo que esta significa y cómo se vincula a otros posibles resultados o datos. La causa radica en la ausencia de metadatos semánticos y ontológicos que proporcionen información adicional para comprender contenidos, significados y la relación entre los datos. Lo que a su vez facilitaría la comunicación y el intercambio de información entre diferentes entidades o sistemas, por ello tienen un comportamiento autónomo completamente reducido (Castells, 2005).

En este sentido, un esquema de operación web podría estar compuesto por agentes software dotados de cualidades tales como autonomía para interactuar con el entorno y tomar decisiones; adaptabilidad para aprender las preferencias del usuario, de otros agentes y de los orígenes de la información, y comunicación para inferir las necesidades y objetivos que tiene el usuario (Peis et al., 2003). Así tendría la capacidad de navegar y ejecutar acciones por sí mismo, con el objetivo de ahorrarle tiempo al usuario y mejorar los resultados. Para ello, es necesario que cada recurso web cuente con una descripción que brinde representación procesable y entendible por programas que puedan asistir, e incluso reemplazar, a los usuarios en tareas rutinarias o de difícil ejecución (Lago y Cacheiro, 2008).

Este es el sentido de Semantic Web, un proyecto que pretende que las páginas en internet contengan algo más que datos para que interpretan los computadores. Es el caso de una página codificada a través de HTML, cuyas *etiquetas* le indica al computador cómo mostrar el conjunto de caracteres en pantalla. Debido a esto, el creador de internet, Tim Berners-Lee, le apuesta a que los computadores puedan interpretar la información que contiene la web y a que sean capaces de tomar decisiones antes de mostrársela al usuario. De acuerdo con este visionario, lo que se espera es que en el futuro un usuario de Semantic Web pueda indicarle a su computador lo que necesita y que éste, al conocer sus preferencias, se encargará de llevar a cabo la tarea (Codina, 2003). La World Wide Web está creciendo rápidamente y a los usuarios les resulta cada vez más difícil obtener información útil y de manera eficiente. Debido a esto, esta iniciativa puede ayudar a reemplazar los denominados *enlaces web* por una *red de significados*, lo que hará que las búsquedas proporcionen contenido útil y de forma más eficiente (Fensel et al., 2003).

Para los usuarios, esta web representa que la información se puede acceder de manera más eficaz, porque actualmente se navega en un nivel bajo. Esto quiere decir que solamente es necesario señalar los enlaces y realizar búsquedas mediante palabras claves; pero, generalmente, estas acciones ofrecen información irrelevante. Un tipo de búsqueda diferente podría encontrar información que utiliza las palabras en contextos diferentes al que se pretende o que descubra enlaces que la utilicen y el abanico de respuestas será mayor. El manejo de ontologías ayuda a resolver este tipo de problemas, porque se centran en el contenido real de una página y de todas las relaciones que pueda tener en la red (Fensel et al., 2003). El modelo Semántico busca crear una web enlazada, que permita localizar, compartir o integrar información con mayor facilidad; incorporando metadatos semánticos y ontológicos que describan el contenido, el significado y las relaciones entre los datos. De esta manera es posible que los agentes puedan evaluar automáticamente la información y entregar resultados más precisos, mejorando el comportamiento autónomo que tienen hasta ahora (Castells, 2005). Los componentes más representativos de este modelo son los Metalenguajes y los estándares de representación XML, RDF, OWL, XML Schema y RDF Schema.

Para lograrlo, se requiere una infraestructura compuesta por un nuevo lenguaje de codificación de páginas, un nuevo lenguaje de marcado, gestión metadatos² y nuevos programas y métodos software para representar el conocimiento, los cuales requerirán de ontologías como procedimientos normalizados para representarlo (Codina, 2003). Además, es necesario que los propietarios de la información, las páginas y demás recursos alojados en la web, utilicen metadatos con el fin de que los motores de búsqueda y el software especializado los puedan encontrar para tomar decisiones. Algunos autores ya han propuesto modelos de arquitecturas por capas para Semantic Web, tal como se observa en la Figura 2.

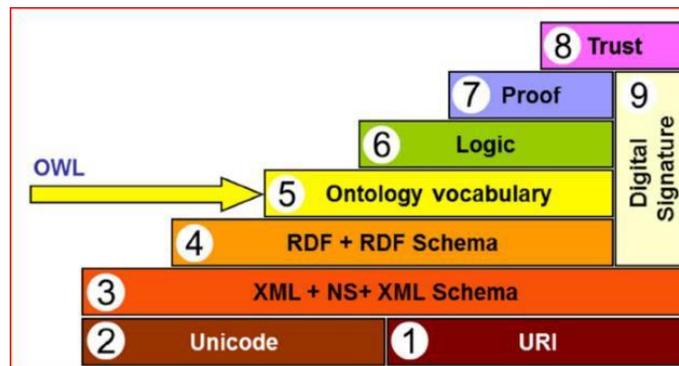


Figura 2. Capas de la arquitectura de Semantic Web (Berners, Hendler y Lassila, 2001)

- Universal Resource Identifier URI. Una cadena formateada que sirve como medio para identificar la ubicación de recursos físicos o abstractos.
- Unicode. Provee un único número para los caracteres, independientemente de la plataforma, programa o lenguaje subyacente. Es requerido para estándares tales como XML, Java, LDAP, WML, etc.
- XML + NS + XMLS. El lenguaje XML permite separar el contenido de la presentación de los documentos a través de la estandarización de formatos de datos, y proporciona la sintaxis superficial para los documentos estructurados. XMLS es un lenguaje que permite definir la estructura de los documentos XML.
- RDF + RDFS. RDF provee un modelo para describir recursos en la web, que son identificables únicamente a través de una URI y que describen la información de manera que pueda ser entendida por las aplicaciones de los computadores. RDFS es el vocabulario que permite describir las propiedades y las clases de los recursos RDF; adicionalmente, aporta la semántica para establecer jerarquías de generalización entre las propiedades y clases. Actualmente, RDF es el modelo más extendido en la comunidad de Semantic Web. Su elemento básico de construcción es la *Sentencia*, que está constituida por dos nodos (*Sujeto* y *Objeto*), que se unen por medio de un arco llamado *Predicado*. En este modelo, los Nodos representan recursos y los Arcos propiedades (Castells, 2005).
- Ontology Vocabulary. OWL es un lenguaje para construir ontologías a través de la descripción detallada de propiedades y clases, tales como Relaciones entre clases, cardinalidad, igualdad, tipologías de propiedades más complejas y caracterización de propiedades o clases enumeradas.
- Logic, Proof y Trust. Son capas aun no exploradas significativamente.
- Digital Signature. Las capas 6, 7 y 8 utilizan firmas digitales como componentes de seguridad.

² Muchas páginas actuales no tienen semántica, sintaxis, ni normas que agrupe las diversas plataformas de metadatos. Para ello se utiliza la norma RDF.

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

Por su parte, Cardoso y Pinto (2015) describen la dirección de las investigaciones futuras con respecto a Semantic Web (Web 3.0), como un pilar fundamental para el desarrollo de nuevos Sistemas de Información. Afirman que, para lograr, se requiere un lenguaje que permita la descripción de conceptos de un dominio dado y la creación de ontologías, como lo hace OWL. Esto conlleva a que en la web se puedan compartir documentos y datos, y a que la búsqueda de información sea más fácil, reutilizable y confiable. OWL es el lenguaje que pueden utilizar las aplicaciones para mejorar los motores de búsqueda y la gestión del conocimiento.

Codina (2003) afirma que las posibilidades de implementar Semantic Web son reducidas a corto y mediano plazo, pero que actualmente se incrementan las posibilidades debido a que se han desarrollado numerosas aplicaciones sobre estas conceptualizaciones. Entre sus argumentos, y en relación con las ontologías, describe que si la Inteligencia Artificial tiene varias décadas de fracasos, por lo menos en su hipótesis fuerte de lograr que los computadores se acerquen a algo parecido a la función humana de pensar, ¿por qué van a tener éxito ahora así de repente? Todavía no está convencido de que estos desarrollos serán exitosos, pero se muestra esperanzado, siempre y cuando haya un cambio de paradigma en las Ciencias computacionales y en las teorías de la documentación. Aún que sobre esta última, argumenta que los documentalistas desde hace décadas podrían estar construyendo una Semantic Web. La cuestión es que si esto es cierto, ¿por qué no han generado aún el cambio de paradigma que él solicita?

En su libro, Codina, Marcos y Pedraza (2009) describen los tipos de cambios que podría aportar Semantic Web: 1) en relación con la Inteligencia Artificial, donde esperan que las tareas que realizan los humanos exitosamente a través del razonamientos, la hagan los computadores a través de sistemas de metadatos, ontologías y lógica formal; 2) en relación con el procesamiento de la información, es decir, el impacto que puede tener con los Sistemas de Información documental. Estos autores no ven un futuro muy promisorio en el primer aporte, pero sí muchas posibilidades en el segundo.

Rodríguez, Merlino y García (2015) presentan la evolución de los métodos de extracción de conocimiento de relaciones semánticas para la web. Éstos son técnicas mediante las que los procesos automatizables analizan fuentes de información no-estructurada (como textos escritos en lenguaje natural) y extraen el conocimiento embebido, para representarlo de manera estructurada y manipulable en procesos de razonamiento automático. Los autores presentan un análisis comparativo de la evolución de los métodos de extracción del conocimiento creados para la web (ver Figura 3). Su objetivo es demostrar cómo, entre 2005 y 2015, cada nuevo método mejora las técnicas de extracción de sus antecesores. De acuerdo con este aporte, se observa el desarrollo del trabajo en Semantic Web. Esto sustenta lo que previamente argumentaba Codina (2003), en el sentido de que a corto y mediano plazo no se tendrán resultados relevantes. Pero también se observa que el cambio de paradigma que él solicita sí se está dando poco a poco.

Desde las áreas de la educación también se están desarrollando propuestas para Semantic Web, entre las que vale la pena mencionar el trabajo de Villarreal et al. (2015), quienes proponen un modelo de Repositorio Semántico de Objetos de Aprendizaje (RSOA) basado en ontologías para entornos virtuales. Su idea es poder compartir el conocimiento y los contenidos multimedia e instruccionales de los cursos *e-learning*, que tienen la característica de ser reutilizados y referenciados en diferentes contextos. Para el repositorio se haría uso de la arquitectura SOA, dado que su enfoque hace que los contenidos sean accedidos e interpretados por agentes, no solamente por personas, llevando esto a un nivel de interoperabilidad semántica.

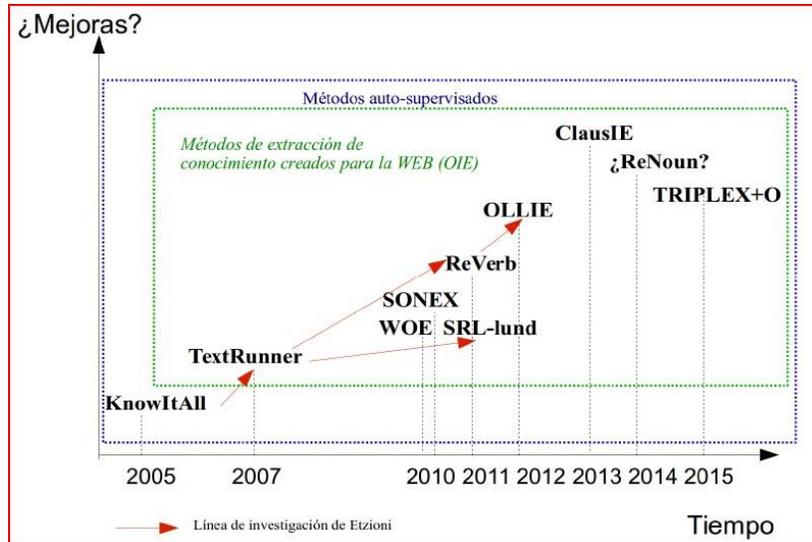


Figura 3. Desarrollo de los métodos de extracción de conocimiento para la web (Rodríguez, Merlinó y García, 2015)

Para Mora et al. (2015) Semantic Web es una herramienta que se puede aprovechar en los procesos docente-aprendizaje para estructurar, almacenar y compartir los contenidos sobre los diferentes cuerpos de conocimiento que se procesan en las instituciones de educación. Para ellos, una de las líneas de investigación docente más activas es desarrollar nuevas formas de gestión de la información en internet. El reto consiste en crear inteligencia colectiva a partir de interacciones web, en lugar de considerar las interacciones de los estudiantes de forma individual. Estas herramientas permiten que los profesores transmitan de mejor manera los conocimientos y que los estudiantes los asimilen más rápido. Así que el cuerpo del conocimiento puede ser explicado desde diferentes puntos de vista obteniendo el mismo resultado.

Otro referente con relación al desarrollo y uso de Semantic Web es el que publican Pandey y Panda (2015) acerca de las diferentes soluciones (herramientas, servicios y proyectos) semánticas desarrolladas, que pueden utilizarse en diferentes áreas del conocimiento, y las soluciones semánticas para bibliotecas digitales basadas en tecnologías web semánticas. Entre las herramientas usadas para la construcción de esta bibliotecas, los autores mencionan: almacenamiento para RDF (Sesame, Jena, YARS), razonadores (KAON, Racer), editores (Protege, SWOOP, MarcOnt Portal) y wikis Semánticos (Makna, SemperWiki). Asimismo, mencionan los desafíos que tiene que afrontar Semantic Web:

- Enfocarse en los motores de búsqueda, ya que hay muy pocos desarrollados de dominio público con enfoques de navegación y búsqueda semántica.
- Dado que la mayoría de las páginas web están desarrolladas en HTML, se requieren estándares semánticos para uso general, que se apliquen a las páginas y así los computadores puedan entender el significado asociado a los conceptos de búsqueda.
- Existen pocos estándares globales y *frameworks* probados para mapeo de dominios y representación del conocimiento, y pocos metadatos en los recursos web.
- Superar la debilidad en la disponibilidad de ontologías específicas de dominio formal, porque son pocas las herramientas desarrolladas para ello. Pero destacan SAPHIRE, un sistema de información en salud basado en semánticas y capaz de rastrear y evaluar situaciones y ocurrencias que podrían afectar la salud pública.

 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

Un valor relevante de las bibliotecas digitales en comparación con la web tradicional, es que en ellas se puede implementar más fácilmente las tecnologías Semantic Web, debido a que las bibliotecas tienen metadatos para catalogar sus recursos. En el área de la Ingeniería del Software, Aguilar y Portilla (2015) proponen un *framework* basado en ODA para describir y componer servicios web semánticos (FODAS-WS), el cual es un marco de referencia que sirve para diseñar y describir servicios Web semánticos automáticamente. El caso de estudio que presentan consiste en el diseño y la implementación de un servicio web semántico, que básicamente es un sistema recomendador, es decir, que ayuda a emparejar a usuarios con productos. Según los autores, este sistema está en capacidad de recomendar dos tipos de productos: objetos de aprendizaje y documentos web, cuyos metadatos se encuentran almacenados en repositorios semánticos de objetos de aprendizaje y en repositorios semánticos de archivos web.

Actualmente, la web es principalmente un sistema donde se almacenan y recuperan datos, sin embargo, el rápido desarrollo, la Sociedad de Conocimiento y la globalización, están exigiendo que se convierta en un sistema para *hacer cosas* y no solamente para *encontrar cosas*. Esto será posible a través de la construcción de un lenguaje semántico, en el que se puedan utilizar reglas acordadas, significados, normas y ontologías. Por lo tanto, los programadores podrán crear sistemas capaces de transmitir un significado, de manera que las máquinas puedan entender lo que se busca. Por ello, Semantic Web es alternativa revolucionaria para los actuales métodos de búsqueda. Aunque es posible que haya algún tipo de resistencia a esta estandarización, es necesaria debido a la cantidad de dispositivos, aplicaciones y lenguajes existentes, y al volumen de datos que existen actualmente. Con esta tecnología se lograrán afianzar los sistemas semánticos, y no hacerlo es arriesgarse a que el Sistema de Información más potente jamás concebido se ahogue en su propia popularidad (Fensel, 2003).

Deep Web

Cuando se investiga el concepto de la Deep WEB, es posible encontrar dos puntos de vista relacionados con su significado: 1) como aquella parte de internet que no es accesible desde los motores de búsqueda tradicionales y que, para lograrlo, se requieren herramientas especializadas. El centro de discusión de este enfoque es su funcionamiento, las herramientas existentes y el potencial como repositorio de datos para que los usuarios resuelvan sus necesidades. 2) Como un medio inseguro, donde es posible alojar y desarrollar actividades que van en contra de la seguridad del mundo.

Pero, ¿qué es y cómo se desarrolla? Deep Web está compuesta por un sin número de bases de datos a las que se puede acceder a través de interfaces de consulta diferentes; la información que se puede obtener de ellas a modo de páginas web se genera de forma dinámica, en respuesta ante una búsqueda; no existen páginas estáticas con vínculos a otros sitios en espera de ser invocados; y, por su estructura, y en algunos casos, es común que los buscadores tradicionales no puedan indexarlas y mantenerlas disponibles. Para lograrlo, sería necesario repensar el modo de ejecución de los mismos y sus *crawlers*, porque debe tener la capacidad de crear con antelación las consultas apropiadas, que arrojen páginas adecuadas y coherentes; dicha tarea requiere comprender el contenido de los sitios consultados para focalizar las consultas (Shestakov, 2008). Deep Web no solamente se debe a la existencia de registros alojados en bases de datos, sitios pornográficos, pedófilos, dictatoriales, expendedores de drogas, prostitución, sicarios y demás, porque aproximadamente el 90% de esta red se alimenta en gran medida de páginas tradicionales y legalmente construidas, aunque no registradas, ya

	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

sea por decisión de su propietario o simplemente por la estructura empleada en su construcción.

Cuándo un usuario realiza una consulta tradicional a través de un buscador de internet, lo que hace el motor de búsqueda es tomar las palabras digitadas y recuperar aquellas páginas que se hallan indexadas a la fecha en su base de datos. La búsqueda traerá como resultado un porcentaje determinado que concuerde con lo buscado. Pero, con seguridad, el motor recupera estas páginas de la web superficial, es decir, aquella a la que tiene acceso la gran mayoría de personas en el mundo. Para hacerse a una imagen de esta situación, Google, aunque ha mejorado su funcionalidad utilizando un sistema de indexado y *crawling* basado en la popularidad de las páginas o documentos (aquellas más visitadas tendrán prioridad para salir en los primeros lugares del listado de resultados), solamente accede al 16% del total de páginas en la web superficial (NEC Laboratories America: <http://www.nec-labs.com/>). Esto quiere decir que los usuarios solamente encuentran el 0.03% de las páginas disponibles, porque el porcentaje restante se aloja en Deep Web (Bergman, 2001). Entonces, se puede decir que esta web es donde se alojan bases de datos y páginas no indexadas y no consultadas a través de motores de búsqueda convencionales.

Bergman (2001) concluye que los motores de búsqueda deben evolucionar para poder acceder tanto a la web superficial como a Deep Web, a la vez que deben satisfacer las consultas de los usuarios. El autor afirma que esto se podría lograr a través de tecnologías de consulta dirigida, donde se involucre la *mega* búsqueda de sitios Deep Web y la *meta* búsqueda de la web superficial. Porque actualmente, aunque existen medios especializados para recolectar información específica, no son útiles para encontrar toda la información disponible sobre un tema o área. Según Bergman, esto sugiere una división del mercado de búsqueda de información en internet, es decir, directorios de búsqueda, que ofrezcan información seleccionada y específica desde la web superficial para satisfacer necesidades de búsqueda populares; motores de búsqueda, para consultas de nivel superior en esta web; e *infohubs*, para la información Deep Web, que proporcione respuestas donde es imperativa la comprensión y la calidad.

En cierta forma, esto corresponde a la tecnología propuesta por Brightplanet (<http://www.brightplanet.com/>) y desarrollada hace tiempo para socavar en Deep Web y traer los resultados de vuelta a la web superficial. Es decir, que al identificar donde residen las bases de datos apropiadas, se puede realizar una consulta simultánea en cada una de las fuentes para recolectar únicamente los resultados deseados con una alta precisión. Esta tecnología utiliza múltiples hilos de múltiples fuentes de consulta y documenta las descargas, luego indexa todos los documentos recuperados y, mediante algoritmos específicos, los clasifica por relevancia de uso (Bergman, 2001).

Aplicando esta tecnología, Brightplanet aprovecha Deep Web, y las debilidades de los motores de búsqueda tradicionales, para encontrar los resultados esperados por los usuarios. Debido a que en su proceso de indexación de páginas web, los motores de búsqueda solamente almacenan las palabras mencionadas con más frecuencia, a la vez que su localización y unos pocos metadatos, esta empresa desarrolla herramientas especializadas para clientes que requieren consultar información sobre temas específicos. Al tener todas las palabras contenidas de una página web en un amplio conjunto de sitios, utiliza tecnologías con capacidades analíticas, ayudándoles a sus clientes a visualizar, analizar y crear inteligencia de grandes conjuntos de datos (Pederson, 2013).

Este tipo de empresas ha logrado almacenar con los años grandes repositorios de datos sobre diferentes sitios web, que utilizan para entregar información a sus clientes. En tal sentido, son los clientes quienes definen su objetivo de búsqueda, y es a través de estas herramientas, desarrolladas con tecnología de web superficial, que logran obtener los resultados esperados desde Deep Web. Un ejemplo de esta relación es lo que ha estructurado la plataforma OpenPlanet. Allí se separa la exploración y los componentes analíticos de la recopilación de datos y el análisis. En otras palabras, no interesa de donde provengan los datos, el objetivo es tomar todos los que defina el cliente e iniciar un proceso para generar los resultados esperados por el mismo. En la Figura 4 se observa el proceso de una exploración desde diferentes fuentes: web superficial, Deep Web, archivos de fuentes RSS, blogs, tweets, correos electrónicos, etc.

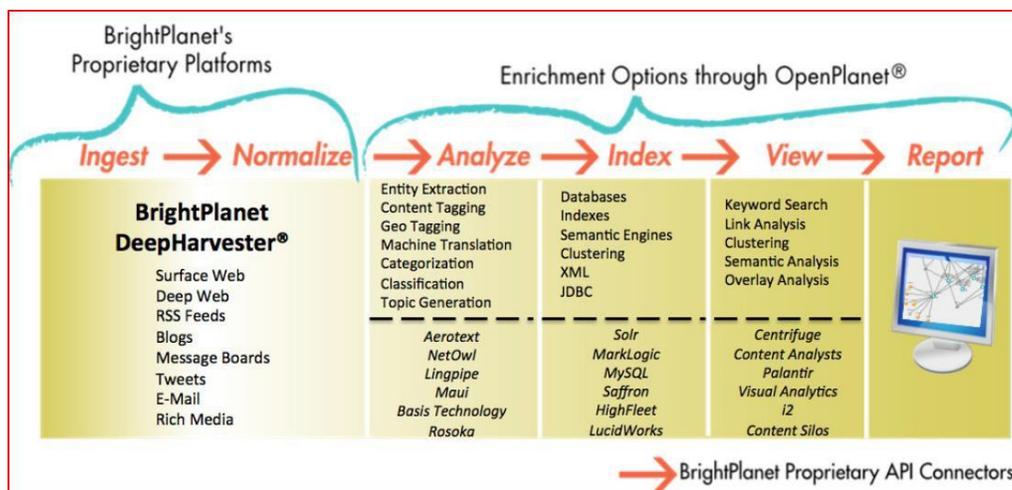


Figura 4. Flujo de trabajo de la plataforma OpenPlanet (Pedersón, 2013)

De forma paralela a estas descripciones y aportes de diversos autores, también es necesario hablar de los aspectos positivos y negativos que se publican alrededor de Deep Web. Aquí es importante destacar que en internet se encuentran muchas y diversas fuentes en las que se hace referencia Deep Web como una web oscura o invisible, pero generalmente son interpretaciones malintencionadas (Pederson, 2013). Claro está, no se puede negar la existencia de peligros que pueden afectar la seguridad de la información y los equipos de los usuarios que navegan por allí, pero lo mismo podría decirse de cuando lo hacen en la web superficial.

En el documental de Mike Radford, *Horizon: Inside the dark Web*, este director describe los pros y los contras de navegar en la web y la forma en que las personas son vigiladas, registradas y analizadas. Desde el surgimiento de internet ha existido una preocupación latente por la privacidad de la información, ya sea personal, gubernamental o empresarial, y que ha sido y es usada por los estados para saber quiénes son los usuarios, qué hacen y si son un riesgo para la sociedad y el mundo. Por otro, también existen personas y organizaciones que utilizan esa información para venderla a otras empresas que necesitan conocer su estilo de vida, qué les gusta y lo que posiblemente podrían comprar. Marc Victoria (2015) hace referencia a una herramienta desarrollada por Financial Times, que puede calcular cuánto podrían pagar ciertos grupos interesados en la información que circula en internet.

Por su parte, los gobiernos le han apostado al tema de salvaguardar la privacidad de su información, con el fin de defenderse de posibles ataques terroristas u otras actividades que afecten a la sociedad. La cuestión que es necesario discutir es qué tanta privacidad se puede perder al navegar en la web. Deep Web se diseñó con el objeto de mantener el anonimato, lo

que puede verse como positivo para mantener la privacidad y tener libertad al momento de adentrarse en internet. Pero, lamentablemente, este tipo de estructuras es visto y aprovechado por otros para desarrollar actividades ilícitas, y es desde este punto de vista supone una amenaza para los estados.

CONCLUSIONES

La mayoría de autores consultados afirman que Semantic Web es un paso evolutivo de la red actual y que su masificación generará avances en relación con los modos de acceso y representación del conocimiento. Pero, para lograr un impacto real se necesita un cambio en la cultura del diseño y desarrollo de páginas web, porque actualmente los sitios alojados en internet carecen de la información semántica requerida. Las páginas deben contener metadatos que les permitan a las herramientas analíticas actuar de manera inteligente y obtener los resultados que esperan los usuarios al realizar las consultas. Por otro lado, algunos de estos autores se muestran escépticos con respecto a un avance rápido de Semantic Web, sobre todo los relacionados con la documentación. Ellos la comparan con el desarrollo de la Inteligencia Artificial y afirman que si después de tantos años de desarrollo no ha podido lograr el avance esperado, esta web tendría dificultades semejantes. Esto se podría evitar con un cambio de paradigma en relación con la estructuración semántica que requiere internet.

Otro aspecto a resaltar es que el progreso de Semantic Web es lento debido a que la mayoría de esfuerzos son independientes. Por ello se requiere de un cambio en la legislación, donde se estandarice el uso de la semántica en todos los sitios web. El desarrollo se visualiza en intentos por estructurar sitios aislados, pero no tanto en motores de búsqueda.

Por otro lado, además que se debe buscar el mejoramiento en el desarrollo de Semantic Web, es claro que Deep Web tiene un alto potencial debido a que en ella se aloja más de las tres cuartas partes de los recursos de información existentes en internet. Pero el caso de ambas es similar, en el sentido de que los esfuerzos por desarrollarlas son mayormente independientes, y se requiere un trabajo más de comunidad.

Aunque los motores de búsqueda tradicionales han progresado en el desarrollo de funcionalidades, todavía no son la herramienta que requieren los usuarios en Deep Web, por lo que se necesita mayor trabajo para alcanzar este territorio y brindar resultados más adecuados en cada consulta de los usuarios. La idea es que no tengan que depender de herramientas especializadas para acceder el volumen de información que está depositada en ella.

Actualmente, las probabilidades de encontrar información de calidad en internet son tres veces más probables en Deep Web que en la web superficial, por lo que el reto para la comunidad es buscar formas de masificar el acceso a ella.

Desde el punto de vista de la seguridad, Deep Web tiene una arquitectura en la que se puede navegar de forma anónima. Esto es útil para salvaguardar la privacidad y transferencia de información de los usuarios, debido a que tiene plataformas y/o redes creadas para tal fin, evitando que se puedan vulnerar sus derechos. La otra cara de la moneda en relación con esta privacidad, es que se presta para realizar actividades ilícitas, lo que dificulta el trabajo de proteger la información de los usuarios.

Como se observa en este artículo, tanto Semantic Web como Deep Web tienen una realidad diferente. La primera es un desarrollo que intenta aprovechar la información a la que los buscadores no pueden llegar, de tal forma que sean los sitios y los computadores los que

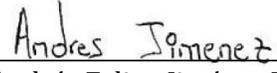
 Institución Universitaria	INFORME FINAL DE TRABAJO DE GRADO	Código	FDE 089
		Versión	03
		Fecha	2015-01-27

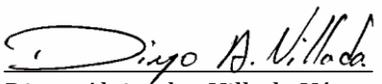
dialoguen para responder las consultas de los usuarios. De esta forma busca entregar los mejores resultados en cada búsqueda que realicen. Mientras que la segunda, con todo el volumen de información que posee y con todo el potencial de aprovechamiento, se encuentra inmersa en mitos y leyendas que no han posibilitado su masificación. Actualmente está restringida a comunidades seleccionadas, debido en parte a que los buscadores tradicionales no la acceden y a que el anonimato en la navegación no es visto desde el punto de vista de la seguridad de los estados y organizaciones.

REFERENCIAS

- Aguiar, C. y da Costa, K. (2013). O que a Deep Web pode oferecer além da Surface Web. Proceedings XV Semana de Informática e XII Mostra de Trabalhos de Iniciação Científica (pp. 1-5). Paranavaí, Brazil.
- Aguilar, J. y Portilla, O. (2015). Framework basado en ODA para la descripción y composición de servicios web semánticos (FODAS-WS). Latin American Journal of Computing 2(2), pp. 15-24.
- Bergman, M. (2001). [White Paper: The Deep Web: Surfacing hidden value](#). The journal of electronic publishing 7(1), online [Aug. 2015].
- Berners, T., Hendler, J. y Lassila, O. (2001). The semantic web. Scientific American, 279(5), pp. 34-43.
- Calaresu, M. y Shiri, A. (2015). Understanding Semantic Web: A conceptual model. Library Review 64(1/2), pp. 82-100.
- Cardoso, J. y Pinto, A. (2015). The Web Ontology Language (OWL) and its Applications. Encyclopedia of Information Science and Technology.
- Castells, P. (2005). La web semántica. En Bravo, C y Redondo, M. (Eds.), Sistemas interactivos y colaborativos en la web (pp. 195-212). España: Ediciones de la Universidad de Castilla-La Mancha.
- Ciancaglini, V. et al. (2014). [Deepweb and Cybercrime - It's not all about TOR](#). A Trend Micro research paper. Online [Aug. 2015].
- Clarke, I. (1999). A Distributed, decentralised information storage and retrieval system. Division of Informatics. Edinburgh University.
- Codina, L. (2003). La Web semántica: Una visión crítica. El profesional de la información 12(2), pp. 149-152.
- Codina, L., Marcos, M. y Pedraza, R (2009). Web Semántica y Sistemas de Información documental. España: Trea.
- Fensel, D. et al. (2003). Spinning the Semantic Web. USA: MIT Press.
- Lago, B. y Cacheiro, M. (2008). Web Semántica Web 3.0. En Marchessou, F. (Ed.), La Web 2.0 y el mundo educativo. XIII congreso Internacional de Informática educativa. Madrid, España.
- Mora, H. et al. (2015). La Web Semántica como herramienta para el apoyo a la docencia. Memorias XIII Jornadas de Redes de Investigación en Docencia Universitaria (pp. 1-14). Alicante, España.
- NEC Laboratories America. <http://www.nec-labs.com/>
- Pandey, S. y Panda, K. (2015). Semantic solutions for the digital libraries based on semantic Web technologies. Annals of Library and Information Studies 61(4), pp. 286-293.
- Parundekar, R., Knoblock, C. y Ambite, J. (2010). Linking the Deep Web to the Linked DataWeb
- Pederson, S. (2013). Understanding the Deep Web in 10 minutes. White paper: Brightplanet.
- Peis, E. et al. (2003). Análisis de la web semántica: Estado actual y requisitos futuros. El profesional de la información 12(5), pp. 368-374.
- Radford, M. (Productor). (2014). Horizon: Inside the dark Web. BBC. London, UK.

- Rodríguez, J., Merlino, H. y García, R. (2015). Revisión sistemática comparativa de evolución de métodos de extracción de conocimiento para la Web. Memorias XXI Congreso Argentino de Ciencias de la Computación (pp. 1-10). Junín, Argentina.
- Shestakov, D. (2008). Search interfaces on the web: Querying and characterizing. Tesis doctoral. Universidad de Turku, Finlandia.
- Varde, A, et al. (2009). Knowledge discovery over the Deep Web, Semantic Web and XML. Lecture Notes in Computer Science 5463, pp. 784-788.
- Victoria, M. (2015). Deep Info! Buscador de datos personales en la Deep Web. Trabajo de fin de grado. Escola Superior Politècnica UPF, Barcelona.
- Villarreal, Y. et al. (2015). RSOA basado en ontologías para entornos virtuales en la educación superior panameña. Virtual educa. Guadalajara, México.
- Zaino, J. (2009). [Why the Deep Web needs the Semantic Web](#). Semanticweb.com. Online [Sept. 2015].
- Zillman, M. (2014). [Deep Web research and discovery resources 2015](#). LLRX.com. Online [Sept. 2015].


Andrés Felipe Jiménez Vélez
Estudiante


Diego Alejandro Villada Vásquez
Estudiante


Prtof. Edgar Serna M.
Asesor

FECHA ENTREGA: _____

Comité trabajo de grado de la facultad

ACEPTADO____ RECHAZADO____ ACEPTADO CON MODIFICACIONES _____

ACTA NO. _____

FECHA ENTREGA: _____

Consejo de Facultad

ACTA NO. _____

FECHA ENTREGA: _____