Construcción de ontologías a partir de tesauros

Luis Manuel Vilches-Blázquez[1], Andrés García Silva[2], Boris Villazón Terrazas[3]
Ontology Engineering Group. Departamento de Inteligencia Artificial.
Facultad de Informática. Universidad Politécnica de Madrid.
{[1]lmvilches, [2]hagarcia, [3]bvilla}@delicias.dia.fi.upm.es

Resumen

Tradicionalmente, los tesauros han sido una de las formas más extendidas para la organización y formalización del conocimiento. Estos, a través de su vocabulario controlado y relaciones, resultan destacados instrumentos para la organización y gestión del conocimiento de un área específica. La importancia de estas formas de organización originó tres estándares para llevar a cabo un proceso de construcción normalizado.

El surgimiento de la Web Semántica permite que los datos sean compartidos y reutilizados a través de diferentes aplicaciones y comunidades. Este hecho conlleva un replanteamiento de las formas de organización del conocimiento y, por tanto, un cambio de estrategia. Estos cambios están vinculados a la necesidad de especificar de manera formal y explícita la semántica asociada a la información de una manera más eficiente que la realizada hasta el momento por los tesauros. Ante esta situación, el uso y desarrollo de ontologías se manifiesta como la mejor forma de especificar la semántica según lo requiere la Web Semántica. Esto está motivando que el proceso de reingeniería y/o la migración de los tesauros tradicionales a ontologías se esté convirtiendo en una tendencia actual.

Palabras claves: Tesauro, Web Semántica, Ontología, SKOS

1. INTRODUCCIÓN

Los tesauros son una de las formas más comunes y tradicionales de organizar y formalizar la información en torno a vocabularios consensuados en dominios concretos. Esto ha provocado que estas formas de organización sean utilizadas en los procesos de recuperación de información con el fin de facilitar a las comunidades de expertos la indexación, consulta y recuperación. Esta razón hace que los tesauros tengan una importancia estratégica en la creación, mantenimiento y reutilización de las fuentes de información asociadas.

El establecimiento de la Web Semántica como un marco común, para permitir que los datos sean compartidos y reutilizados a través de diferentes aplicaciones y comunidades, conlleva un cambio de estrategia en la

organización y gestión de las fuentes de información contenidas en la Web. Esta situación exige el uso de instrumentos que permitan hacer explícito y formal el significado de la información de una manera más avanzada que los tradicionales tesauros. Este es el caso de las ontologías, ampliamente utilizadas en el dominio de la Inteligencia Artificial, son los componentes que permiten especificar la semántica de la información.

El proceso de reingeniería y/o la migración de los tesauros tradicionales a ontologías se están convirtiendo en una tendencia actual. Esta tendencia está basada en el hecho de que la estructura de los tesauros (los términos y sus relaciones) y su uso como vocabularios aceptados por una comunidad de usuarios se adaptan perfectamente a la concepción de las ontologías.

En este capítulo se pretende analizar la utilidad de los tesauros como fuente de datos para la construcción de ontologías en el contexto de la Web Semántica. En la sección 2, se lleva a cabo una introducción a los tesauros, describiendo su estructura y formas de representación. La sección 3 describe, brevemente, algunas características de la Web Semántica y las ontologías. Asimismo, se describen los problemas que presentan los actuales tesauros en relación a la Web Semántica. Las recomendaciones del *W3C* para la definición de tesauros en el contexto de la Web Semántica son presentadas en la sección 4. La sección 5 presenta diferentes técnicas del estado del arte para realizar el proceso de migración desde los clásicos tesauros a ontologías. Por último, en la sección 6 se presentan, de forma breve, algunas conclusiones a este capítulo.

2. TESAUROS

Un tesauro, según *International Standard Organization* (ISO), es un vocabulario de un lenguaje de indexación controlado (es decir, un conjunto controlado de términos extraídos del lenguaje natural y utilizados para representar, de forma breve, los temas de un documento) y organizado formalmente con objeto de hacer explícitas las relaciones, a priori, entre conceptos (por ejemplo, “más genéricos” o “más específicos que”) (ISO 2788, 1986; ISO 5964, 1985).

El establecimiento y desarrollo de tesauros es una tarea compleja. Por esta razón, existen tres estándares internacionales (ISO 5964, 1985; ISO 2788, 1986; ANSI/NISO Z39.19, 1993) que guían los esfuerzos en el control del vocabulario, establecimiento de relaciones entre términos y la representación de términos en un tesauro. Asimismo, la normalización de los métodos para la construcción de un tesauro es considerado un elemento clave para obtener compatibilidad entre los diferentes tesauros. A continuación se van a reflejar las principales características de las actividades del proceso de construcción de un tesauro:

2.1 Control del vocabulario

La utilización de sinónimos permite expresar un mismo significado a través de diferentes conceptos (por ejemplo, agro-industria, industria de productos agro-alimenticios y/o empresas agrícolas). Uno de los propósitos de un tesauro es eliminar los problemas ocasionados por los sinónimos. Así, un sinónimo es elegido, de manera más o menos arbitraria, como un descriptor o término preferente (ISO 2788, 1986) (palabras o expresiones que son utilizadas sistemáticamente en la indexación para representar un concepto específico). El resto de sinónimos para el mismo concepto se les asigna el estatus de no descriptores o términos no preferentes (ISO 2788, 1986) (sinónimos o cuasi-sinónimos con respecto a un término preferente que proporciona un punto de entrada desde el que el usuario es guiado por unas instrucciones).

Por otro lado, un término puede tener varios significados (por ejemplo, “prensa”, puede estar relacionado con la industria de los periódicos, periodistas, máquinas de impresión o diferentes herramientas mecánicas). Por este motivo, otro de los propósitos de un tesauro es eliminar los problemas relacionados con la homonimia. De esta manera, cada término preferente es puesto en contexto de forma que su significado no resulte ambiguo.

2.2 Relaciones

Las relaciones son otro de los componentes esenciales a ser considerados en el establecimiento y desarrollo de un tesauro. Además, estas hacen posible visualizar y separar fácilmente las conexiones entre cada término contenido. Por esta razón, en (ISO 5964, 1985; ISO 2788, 1986; ANSI/NISO Z39.19, 1993) se definen las posibles relaciones que pueden ser utilizadas entre los términos en tesauros monolingües y multilingües.

Relaciones de equivalencia. Este tipo de relaciones es establecida entre los términos preferentes y no preferentes cuando un término o más de uno es referido al mismo concepto, formando una equivalencia completa.

Relaciones de jerarquía. En esta relación un término de categoría superior representa una clase y los términos subordinados corresponden con sus miembros o partes. A este nivel existen tres subtipos de relaciones lógicas:

- Relaciones genéricas. Identifican la relación dentro de una clase o categoría y sus miembros o partes.
- Relaciones jerárquicas parte-todo. El nombre del todo es utilizado como un término superior y el nombre de las partes como un término subordinado que involucran el nombre del todo al que pertenece.

- Relaciones de enumeración. Identifican la existencia de relación dentro de una categoría general de objetos o sucesos, expresados con un nombre común y un caso individual de esta categoría.

Relaciones de asociación. Se ocupan de aquellas relaciones que son establecidas entre los términos no equivalentes. A pesar de los términos asociados están relacionados mentalmente, no pueden ser conectados jerárquicamente. Los términos relacionados deben estar necesariamente sujetos a un control, es decir, uno de estos términos tiene que ser incluido en cualquier otra definición de términos.

Estos tipos de relaciones en un tesauro son expresadas con un conjunto de etiquetas. Estas etiquetas preceden a los términos y tienden a dar cierta estructura y clarificar las relaciones entre los términos de un área específica de conocimiento. Algunas de estas etiquetas son: “Término Genérico (TG)”, “Termino Específico (TE)”, “Término Relacionado (TR)” “Use (USE)”, “Use por (UP)”, etc.

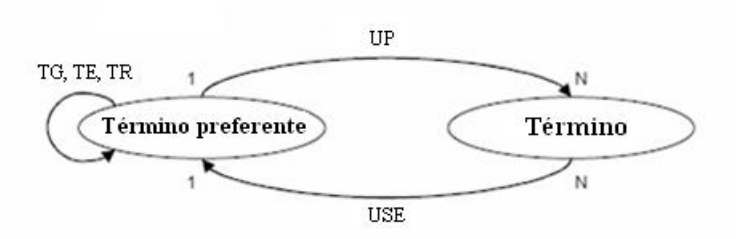

Figura 1. Relaciones básicas del tesauro, conforme a ISO 2788: 1986. (van Assem et al., 2004)

Los tres tipos de relaciones básicas entre términos, comentadas con anterioridad, son utilizadas para relacionar términos preferentes y no preferentes como puede apreciarse en la Figura 1. Los términos preferentes están relacionados con términos no preferentes o no descriptores a través de la relación “Use por (UP)”. USE es la inversa de esta relación. Sólo los términos preferentes tienen relaciones del tipo “Término Específico”, “Término Genérico” y “Término Relacionado”. Las relaciones de “Nota de Alcance (NA)”, como puede apreciarse en Tabla 1, son utilizadas para proporcionar una definición de un término.

2.3 Representación

Tras la consideración del proceso de control de vocabulario y el establecimiento de relaciones, el siguiente paso concierne a la selección del mejor esquema de presentación de un tesauro. Los tesauros existentes,

normalmente, pertenecen a uno de los siguientes tipos básicos de representación: alfabética, sistemática o gráfica (ISO 2788, 1986; ISO 5964, 1985). Algunos de los tesauros existentes pueden incluir elementos de varios tipos de presentación.

Probablemente, la **presentación alfabética** es la forma más simple en lo que respecta a la construcción y reproducción. Los términos preferentes y no preferentes son organizados en una única secuencia alfabética.

En la **presentación sistemática** el tesauro está dividido en dos secciones;

- Categorías de términos o jerarquías. Organizadas según sus significados e interrelación lógica. Esta parte se denomina sección sistemática.
- Índice alfabético. Utilizado para guiar a los usuarios en las respectivas partes de la sección sistemática.

La **presentación gráfica**, los términos indexados (con frecuencia, términos preferentes) son visualizados en un diagrama (un esquema semántico), organizado de tal manera que un usuario puede tener una visión global de todos los términos y sus relaciones. Dos de las formas más comunes de representación son las estructuras en árbol y los gráficos de flechas. Junto a estas representaciones, normalmente, se acompaña un índice alfabético que guía a los usuarios a las diferentes partes del diagrama.

3. TESAUROS Y WEB SEMÁNTICA

No hay duda de que la Web provee una plataforma de acceso global a la información. Sin embargo, hay un número importante de cuestiones que deben tenerse en cuenta para que su potencial sea totalmente desarrollado. La Web no fue inicialmente concebida como una herramienta para acceso global de información, por lo cual los estándares subyacentes para la administración de la información no son totalmente adecuados. Debido a la naturaleza de la arquitectura de Internet, la información de temas similares está esparcida a través de diferentes servidores alrededor del mundo y, hoy en día, sólo se cuenta con unas pocas herramientas para integrar información relacionada de diferentes fuentes. Por lo cual usualmente es muy difícil encontrar recursos en la Web.

Los tesauros se han utilizado como herramientas útiles para los procesos de recuperación de información al proveer vocabularios controlados y formalmente organizados, clarificando las relaciones jerárquicas entre términos, aunque no su significado (ISO 5964, 1985; ISO 2788, 1986). En el contexto tecnológico actual, los tesauros no son suficientes debido a que si se quiere crear bases de conocimiento, tales como las requeridas por la Web Semántica, éstos sólo proveen una parte de las funcionalidades requeridas (Wielinga et al., 2001). Sin

embargo, los tesauros constituyen una fuente importante de información estructurada y consensuada para la construcción de ontologías que como se expone más adelante son un pilar fundamental para la Web Semántica.

La Web Semántica (Berners-Lee et al., 2001) (Berners-Lee, 1998) representa la próxima iteración de la *World Wide Web*, en la cual los ordenadores conectados en red serán capaces de extraer significado unos de otros en lugar de simples cadenas de caracteres. Esencialmente, la Web Semántica servirá como un medio por el cual los ordenadores puedan localizar, organizar y desplegar, de manera más efectiva, información a través de las fronteras entre aplicaciones, empresas y comunidades.

Los lenguajes de la Web Semántica como RDF (*Resource Description Framework*), RDFS (*Resource Description Framework Schema*) y OWL (*Ontology Web Language*) proveen un formalismo de datos para describir recursos, sus propiedades, interrelaciones y categorías. Estos lenguajes permiten que la información sea procesable automáticamente por los ordenadores al formalizar y hacer explícito su significado. La responsabilidad de asignar metadatos significativos se deja a cargo de los usuarios. De esta manera, la Web Semántica simplemente proveerá un marco de trabajo estandarizado para habilitar la integración e intercambio de esta información (Wilcox et al., 2005).

Las ontologías son los principales instrumentos de la Web Semántica para superar los problemas de administración de la información de la Web actual. Las ontologías llevan la conceptualización un paso adelante, reestructurando formalmente los términos y proporcionando relaciones más elaboradas entre conceptos, en comparación con los tesauros tradicionales. Los usuarios tendrán a su disposición un contexto más amplio para evaluar la utilidad de la información al tener disponible la definición formal de los términos, y una descripción precisa de las relaciones entre ellos. De la misma manera, con la estructura formal del contexto y el significado de los términos, las ontologías se convierten en una pieza importante de la Web Semántica, cuando ésta se describe como "*una extensión de la Web actual, en la cual la información recibe un significado bien definido, que ayuda a los ordenadores y a las personas a trabajar mejor juntos*" (Berners-Lee et al., 2001).

3.1 Problemas de los Tesauros relativos a la Web Semántica

En el dominio geográfico, existen diferentes tesauros bien establecidos y autorizados por diversas organizaciones, una descripción detallada puede encontrarse en (Vilches-Blázquez et al., 2007). Estos tesauros proveen un conjunto de términos jerárquicamente estructurados acerca de los cuales una

comunidad de usuarios ha alcanzado un acuerdo. Este es precisamente el tipo de conocimiento previo requerido en las aplicaciones de la Web Semántica. Sin embargo, en el proceso de construcción de tesauros hay varios problemas que constituyen una dificultad para lograr alcanzar los requerimientos de interoperabilidad semántica que necesita el nuevo contexto Web. Los problemas se listan a continuación:

Interpretación no ambigua: El primer requerimiento para un tesauro es proveer una estructura jerárquica que tenga una interpretación sin ambigüedad. Sin embargo, la herencia simple típica de los tesauros limita la cantidad de información que puede ser derivada de su posición en la jerarquía acerca de un término. Este problema se aborda por medio de la clarificación de términos, de tal manera que puedan ser incluidos en un contexto o en otro, dependiendo de la explicación relacionada al término. Un concepto como “paisaje” puede representarse por dos términos diferentes: Paisaje (ambiente) y Paisaje (representación). Esta clase de distinciones del mismo concepto puede no ser clara para los usuarios, además es difícil decidir dónde ubicar las subclases del concepto (Wielinga et al., 2001).

Categoría única de términos: Otro problema a considerar en los tesauros actuales es la ubicación o inclusión de términos bajo una cierta categoría de clasificación teniendo en cuenta que en realidad un término puede estar incluido en otras categorías. Esto es un problema no sólo cuando el usuario debe seleccionar una categoría de una jerarquía en busca de un término, sino también para los procesos de búsqueda que usan herencia (Wielinga et al., 2001). La búsqueda de conocimiento acerca de ciertas características geográficas, en relación con otras características o áreas, requiere una interrelación sobre diferentes partes de la jerarquía de un tesauro. Vinculado con este problema está el hecho que las relaciones no tienen una clara precisión semántica. Este es el caso de los términos “más genérico”, “más específico”, “use por”, “relacionado” y “equivalente”. La imprecisión afecta directamente la ubicación del término en la jerarquía que a su vez limita la recuperación de información a partir del término. Esta situación evidencia que la semántica de un tesauro, aun siguiendo las recomendaciones del estándar ISO, no es tan precisa como la definida por los lenguajes de la Web Semántica (Wilson et al., 2002)

Términos preferentes: Un problema común en la creación de los tesauros es que sus desarrolladores deben enfrentarse con un gran conjunto de términos posibles de los que deben escoger aquellos que representen el concepto en cuestión (Wielinga et al., 2001). Esto origina la necesidad de elegir términos preferentes, de acuerdo a la frecuencia de uso del término en el área de conocimiento en particular y a la vez relacionarlos con términos no preferentes. Una solución para este problema es restringir los conjuntos de términos para cada campo en particular, basados en una descripción parcial. Además, en

algunos casos, varios campos pueden ser inferidos de la información disponible en otros campos (Wielinga et al., 2001), aunque la inferencia en tesauros es muy limitada debido a la naturaleza básica de las relaciones y a la ausencia de relaciones ad-hoc.

Formato de archivo de los tesauros: Los formatos de archivos comunes para los tesauros también son un problema importante. Los tesauros usualmente son expresados en formatos nativos, frecuentemente archivos XML propietarios, ASCII o esquemas relacionales. Estos formatos no son compatibles con los estándares de la Web Semántica, es decir, con RDF (van Assem et al., 2004), RDFS (W3C, 2004) y OWL.

De acuerdo a los problemas expuestos anteriormente, se puede decir que los tesauros presentan importantes inconvenientes para ser utilizados directamente en la formalización de la información de acuerdo a los requerimientos de la Web Semántica. Esto hace que el uso de tesauros no sea adecuado para el proceso de recuperación de información de la Web, aun teniendo en cuenta las mejoras sustanciales de las capacidades de las herramientas de búsqueda actuales (Vilches et al., 2006a).

3.2 Tesauros y ontologías

Una definición sencilla de ontologías podría ser: Un tesauro mejorado, cuando se define tesauro como un instrumento que establece el vocabulario de un cierto campo temático y las relaciones entre los términos de dicho campo. Esta definición de tesauro encaja perfectamente con la definición de ontología dada en (Gruber, 1993; Studer et al, 1998), donde los autores afirman que una ontología constituye "*una especificación formal y explicita de una conceptualización compartida* (Gruber, 1993)*, donde la semántica de la información se hace explicita por medio de los objetos, sus relaciones y las propiedades que los caracterizan, en un lenguaje formal que sea entendible por los ordenadores*" (Studer et al, 1998).

Siguiendo la línea marcada por Gruber, muchas estructuras actuales de representación de conocimiento pueden clasificarse como ontologías. En (Lassila et al., 2001) se identifica el espectro ontológico presentado en la Figura 2**¡Error! No se encuentra el origen de la referencia.**. Esta representación lineal del espectro inicia en la parte izquierda con estructuras muy simples referentes a vocabularios controlados (catálogos y glosarios) donde el significado es definido por medio del lenguaje natural. Un nivel más adelante se ubica a los tesauros que proveen semántica por medio de las relaciones entre los términos, aunque estas relaciones no definen explícitamente una jerarquía. Para algunos investigadores las ontologías deben proveer una jerarquía explicita para ser

considerada propiamente como una ontología. Es por este motivo que se dibuja una línea diagonal que atraviesa el espectro dividiendo las estructuras que proveen una jerarquía explicita de las que no la proveen.

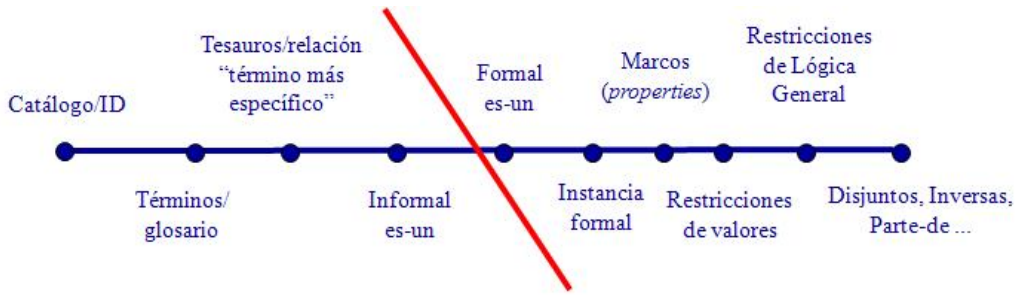

Figura 2. Espectro ontológico

Algunos trabajos de investigación se han enfocado en el proceso de conversión de tesauros existentes a RDF(S) y OWL (Wilson, 2002; Hall, 2001; van Assem et al., 2004; Cross et al., 2000; Wielinga et al., 2001). En el campo de la Ingeniería Ontológica (Gómez-Pérez et al., 2003) está ampliamente reconocido que un tesauro puede ser referido como una ontología. Aunque debido a sus limitaciones los tesauros caen en la categoría de las denominadas ontologías ligeras (Corcho et al., 2003). Estas ontologías ligeras incluyen conceptos, taxonomías de conceptos, relaciones entre los conceptos y propiedades que describen los conceptos.

En la actualidad para que un tesauro pueda ser usado en la Web Semántica, debe ser formalizado por medio de un lenguaje propio de la Web Semántica. Esta formalización puede realizarse manteniendo las relaciones existentes en el tesauro o pueden ser transformadas en relaciones más propias de las tecnologías de la Web Semántica, como “subclase de” o “parte de”. De acuerdo a (Wielinga et al., 2001), un tesauro debería satisfacer los siguientes criterios para que pueda ser útil en la Web Semántica: (1) debería tener una estructura jerárquica estricta de subclases y superclases, (2) debería basarse en conceptos únicos en lugar de en términos del lenguaje natural y (3) debería ser representado en un formato que cumpla con los estándares de la Web Semántica.

4. SKOS

Simple Knowledge Organisation Systems (SKOS) es una familia de lenguajes formales diseñados para la representación de un tesauro, esquemas de clasificación, taxonomías, sistemas de encabezado de materias o cualquier

otro tipo de vocabulario controlado estructurado. SKOS está construido sobre RDF y RDF(S) y su principal objetivo es permitir la publicación de manera sencilla de vocabularios estructurados controlados para la Web Semántica. En SKOS existen diferentes especificaciones derivadas de la modularización y extensibilidad de la familia de estos lenguajes:

SKOS Core Vocabulary Specification (W3C, 2005a). Es una aplicación de RDF que puede ser utilizada para expresar un esquema de conceptos como un grafo RDF. La utilización de RDF permite a los datos ser relacionados y/o fusionados con otros datos, permitiendo a las fuentes de datos ser distribuidas a través de la Web, para ser compuestos e integrados de forma significativa. Este documento da una visión general de la referencia de estilo del vocabulario *SKOS*. Además, describe las políticas de propiedad, nombrado, persistencia y cambio por cualquier vocabulario gestionado por *SKOS*.

SKOS Core Guide (W3C, 2005b). Este documento es una guía para la utilización de *SKOS Core Vocabulary*, destinado a aquellos usuarios que tengan un entendimiento básico de los conceptos de RDF.

SKOS Primer (W3C, 2008a). Este documento es una guía de usuario para aquellos que pretenden representar su esquema de conceptual mediante la utilización de SKOS. *SKOS Primer* presenta dos niveles:

- En el SKOS básico, los recursos conceptuales (conceptos) son identificados con URIs (*Uniform Resource Identifiers*), etiquetadas con cadenas de caracteres (*strings*) en uno o más lenguajes naturales, documentados con varios tipos de notas, semánticamente relacionados en jerarquías informales y redes de asociación y agregadas en esquemas de conceptos.
- En el SKOS avanzado, los recursos conceptuales pueden ser mapeados a través de esquemas de conceptos y agrupados en colecciones ordenadas o etiquetadas. Las relaciones entre etiquetas de conceptos pueden ser especificadas. Finalmente, el propio vocabulario de SKOS puede ser extendido de forma conveniente a una comunidad particular o combinado con otros vocabularios de modelado.

SKOS Reference (W3C, 2008b). Este documento define una modelo de datos común para compartir y relacionar sistemas de organización de conocimiento vía Web. Asimismo, este documento es la especificación normativa de SKOS, prevista para los usuarios que estén involucrados en el diseño e implementación de sistemas de información que dispongan de un buen conocimiento de las tecnologías de la Web Semántica, especialmente de RDF y OWL.

Quick Guide to Publishing a Thesaurus on the Semantic Web (W3C, 2005c): Este documento describe de forma breve como expresar el contenido y estructura de un tesauro y los metadatos sobre un tesauro, en RDF. El uso de RDF permite a los datos ser relacionados y/o fusionados con otros datos en formato RDF por aplicaciones de la Web Semántica. En la práctica, esto significa que las fuentes de datos pueden ser distribuidas a través de la Web de una manera descentralizada.

A continuación se va a reflejar un ejemplo, extraído de (W3C, 2005c) basado en el *United Kingdom Archival Thesaurus* (UKAT) en el que se mostrarán las principales relaciones establecidas por las diferentes normas para el establecimiento y desarrollo de un tesauro (ISO 5964, 1985; ISO 2788, 1986; ANSI/NISO Z39.19, 1993) y su transformación a SKOS. Este ejemplo es mostrado en la Tabla 1

. La misma información, expresada como un grafo RDF utilizando el *SKOS Core Vocabulary*, se presenta en la Figura 3.

Término:	**Cooperación económica**
Use por:	Co-operación económica
Término Genérico:	Política económica
Término Específico	Integración económica
	Cooperación económica europea
	Cooperación industrial europea
	Cooperación industrial
Término Relacionado:	Interdependencia
Nota de Alcance:	Incluye medidas cooperativas en banca, comercio, industria, etc., a nivel inter e intra países.

Tabla 1. Extracto del *United Kingdom Archival Thesaurus* (W3C, 2005c)

Cada concepto del UKAT tiene una URI asignada. Las URIs son identificadores únicos globales que pueden ser utilizados para referirse a un recursos de manera inequívoca. Cualquier cosa puede ser un recurso, no sólo documentos Web. Las URIs pueden, por tanto, ser utilizadas como identificadores para cualquier elemento. Por ejemplo, la URI http://www.ukat.org.uk/thesaurus/concept/1750 se refiere a un concepto en un tesauro. Las URIs asignadas a los conceptos en un tesauro permiten referirse a ellos de manera inequívoca desde cualquier contexto. Por otro lado, cuando se expresa el contenido de un tesauro, como UKAT, en RDF utilizando *SKOS Core*, cada término preferente llega a ser una propiedad preferente asignada a un

concepto y cada término no preferente se convierte en una propiedad alternativa para un concepto. De esta manera, las propiedades *skos:prefLabel* y *skos:altLabel* permiten la asignación de etiquetas léxicas alternativas y preferentes a un recurso. La

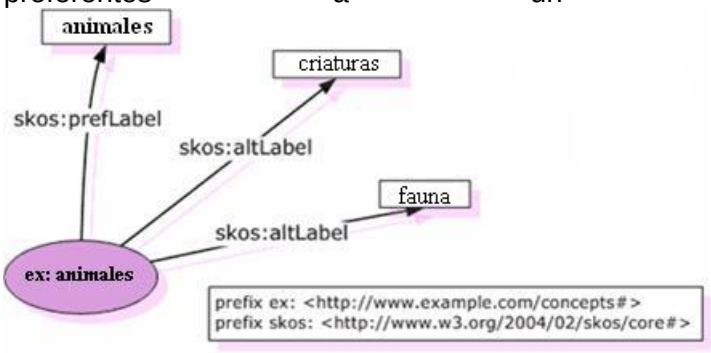

Figura 4 proporciona un ejemplo del uso de estas etiquetas.

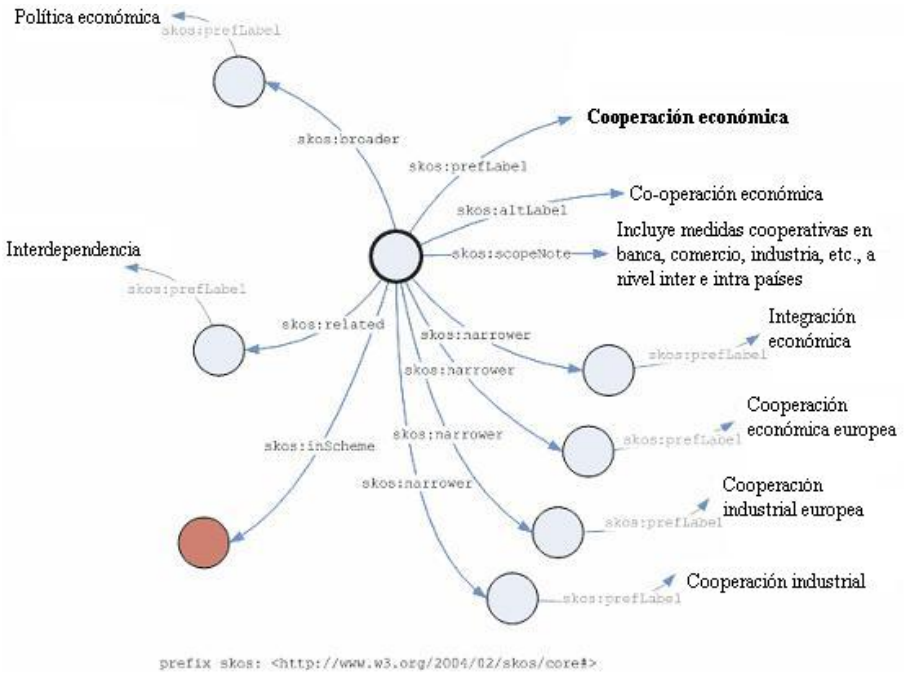

Figura 3. Grafo RDF (W3C, 2005c)

Las propiedades *skos:prefLabel* y *skos:altLabel* son subpropiedades de *rdfs:label*. Las restantes relaciones de los tesauros son establecidas como las relaciones semánticas dentro del *SKOS Core Vocabulary* (W3C, 2005a). Asimismo, se incluyen propiedades con objeto de reafirmar las relaciones semánticas entre conceptos. Así, la propiedad *skos:semanticRelation* comprende

skos:broader (refleja la relación Término Genérico), *skos:narrower* (se refiere a la relación Término Específico) y *skos:related* (corresponde a la relación Término Relacionado).

Las propuestas de las diferentes especificaciones de SKOS reflejan el esfuerzo del W3C por desarrollar estándares para el proceso de conversión de los tesauros tradicionales a ontologías de la Web Semántica. Tales ontologías permitirán la definición autorizada y distribuida de vocabularios que soporten la referencia cruzada. Además, las representaciones de ontologías son planificadas para cubrir el papel actualmente asumido por los tesauros. Este hecho se deriva de que una migración o soporte para su coexistencia se hace necesaria si las ontologías son adoptadas y asimiladas en las infraestructuras de recuperación de información existentes.

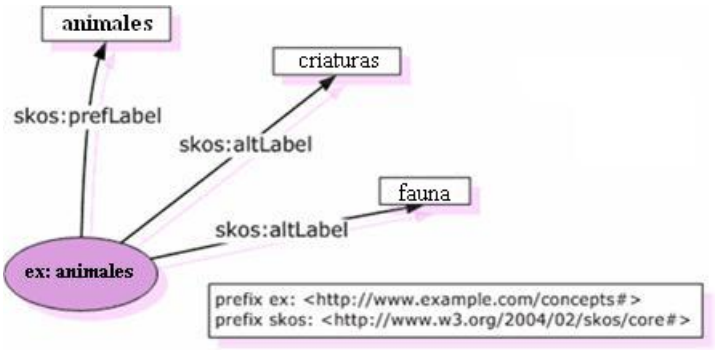

Figura 4. Grafo RDF utilizando etiquetas preferentes y no preferentes para un concepto (W3C, 2005b)

Por otro lado, la conversión de los recursos existentes a RDF implica varios aspectos: (1) la calidad de un tesauro puede ser mejorada utilizando la semántica que aporta RDF(S), ya que los tesauros utilizan relaciones con una semántica no claramente definida o aplicadas de forma ambigua. (2) los tesauros convertidos pueden ser comprobados utilizando razonadores estándar, identificando la pérdida de categorías conceptuales y relaciones inversas, por ejemplo, Término Genérico/Término Específico (van Assem et al., 2004).

Bajo este contexto aparecen diversos enfoques cuyo objetivo consiste en migrar los tesauros clásicos a SKOS. El tesauro AGROVOC (FAO, 1997) y el proyecto *Agricultural Ontology Service* (AOS) son un buen ejemplo. Esta iniciativa pretende contribuir a desarrollar terminologías especializadas de dominio que contribuyan a estructurar y normalizar la terminología agraria en diversos idiomas. La pretensión es que dichas terminologías puedan ser usadas en múltiples sistemas en el ámbito de la agricultura y proporcionar diversos

servicios que mejoraren el soporte de la gestión de información en el entorno Web.

Una de las finalidades clave es añadir más semántica al tesauro AGROVOC para expandir y mejorar la especificación de relaciones entre conceptos. De esta manera, AOS conseguirá una mayor interoperabilidad entre los sistemas que gestionan la información agraria. Por tanto, FAO se ha involucrado en el proceso de reingeniería del tesauro AGROVOC y, especialmente, en el desarrollo de ontologías de dominio específicas tales como seguridad alimenticia, industria pesquera y nutrición alimenticia. Una descripción más detallada de estas diferentes ontologías puede encontrarse en (Vilches-Blázquez et al., 2007).

En resumen, el proceso de rediseño del tesauro AGROVOC y los servicios en un sistema basado en ontologías, utilizando OWL conlleva la conversión a un nuevo sistema, donde destaca el valor añadido de las ontologías sobre los tradicionales tesauros, como consecuencia del enriquecimiento ofrecido por el contexto de la Web Semántica y su énfasis en la interpretación de datos significativos. Las ontologías sirven como el significado de enumeración de todas las posibles relaciones entre elementos que pueden ser utilizados para organizar y clasificar datos. De manera sencilla, las ontologías de la Web Semántica definen explícitamente la estructura y jerarquía de un dominio particular y éstas son capaces de reconocer la conexión entre elementos de datos a pesar del hecho de que los ordenadores no son, hasta el momento, capaces de reconocer los significados verdaderos de estas conexiones (Wilcox et al., 2005).

5. MÉTODOS PARA LA TRANSFORMACIÓN DE TESAUROS A ONTOLOGÍAS

El proceso de construcción de ontologías a partir de tesauros pretende reutilizar recursos disponibles con el objetivo de ahorrar tiempo, esfuerzo y aprovechar el conocimiento disponible en este tipo de fuentes de información. Este hecho está motivando que la comunidad investigadora del contexto de la ingeniería ontológica comience a reutilizar los tesauros que ya poseen un importante grado de consenso. Esta reutilización involucra necesariamente un proceso de transformación y/o reingeniería de los tesauros clásicos. A continuación se describen los principales enfoques para la transformación de tesauros a ontologías.

(van Assem et al., 2004) presentan un método para convertir los tesauros de su formato nativo a RDF(S) y OWL. Este método se encarga de los recursos implementados en (1) un formato de propietario, (2) una base de datos relacional, y (3) ficheros XML. Este método transforma semi-automáticamente la totalidad del contenido del recurso en un esquema de una ontología. En todo el proceso de transformación se han adoptado decisiones con respecto a la sintaxis

y la semántica de la representación resultante. Este método consta de los siguientes pasos:

- *Preparación*. Se analizan las siguientes características del tesauro: (1) modelo conceptual, (2) relación entre el modelo conceptual y modelo digital, (3) relaciones con estándares, y (4) la identificación de cuestiones multilingües.
- *Conversión Sintáctica*. Este paso se centra en los aspectos sintácticos del proceso de conversión de la implementación de la fuente a RDF(S).
- *Conversión Semántica*. En este paso las definiciones de clases y propiedades se complementan con las restricciones de RDF(S) y OWL. La salida de este paso debe ser utilizado en aplicaciones como una interpretación específica del tesauro, no como una conversión estándar.
- Estandarización. Este paso opcional se crea una representación del tesauro en un esquema estándar. Una posible opción es SKOS (W3C, 2005a).

La ontología resultante se expresa en RDF(S)/OWL y el método genera una única ontología.

En (Hahn, 2003, Hahn et al., 2003) estos autores presentan un método que extrae el conocimiento conceptual de un tesauro médico (*Unified Medical Language Systems* - UMLS), y semi-automáticamente convierte este conocimiento en un sistema formal de lógica descriptiva, LOOM. Este método transforma el contenido del recurso en un esquema de la ontología. Este enfoque se compone de los siguientes pasos:

- Generación automática de las expresiones terminológicas. En este paso las relaciones como *parte de*, *es un* o *tiene localización* se tienen en cuenta.
- Comprobación automática de consistencias utilizando el clasificador LOOM. La base de conocimiento sin refinar es inmediatamente verificada por el clasificador para ver si contiene ciclos de definición e inconsistencias.
- Restitución manual de las consistencias. En caso de encontrarse inconsistencias, un experto del dominio se encarga de resolver las inconsistencias. Después, el clasificador se vuelve a ejecutar para comprobar si la modificación de la base de conocimientos es incompatible con los cambios realizados.
- Conservación manual de la base de conocimientos. En este paso se consideran las relaciones que no se toma en cuenta en los pasos anteriores (por ejemplo, *hermanos de* o *asociado con*) son incluidas.

La ontología resultante se expresa en un sistema formal de lógica descriptiva y el método genera sólo una ontología.

(van Assem et al., 2006) presentan un método para convertir los tesauros a SKOS (W3C, 2005a) RDF / OWL esquema, que es una propuesta estándar bajo desarrollo por el *W3Cs Semantic Web Working Group*. El desarrollo de este enfoque se basa en un proceso con los siguientes componentes:

- El objetivo general de este enfoque es apoyar la interoperabilidad de los tesauros codificados en RDF / OWL. Los requisitos son los siguientes: (1) producir programas de conversión que transformen la representación digital de un tesauro a SKOS. El programa de conversión debería producir SKOS RDF. (2) Realizar una conversión completa del tesauro (es decir, la ontología resultante tiene toda la información que está presente en el tesauro original), siempre que con ello no se infrinja el anterior requisito.
- Comparación con los métodos existentes. Aquí los autores comparan los objetivos y los requisitos a los métodos existentes para elegir el más adecuado. Los autores compararon los trabajos de (Soergel et al., 2004; W3C, 2005c; van Assem et al., 2004).
- Desarrollar los pasos del método. Los Autores eligieron el método de Miles et al. (W3C, 2005c), que tiene un objetivo y requisitos comparables, como punto de partida y lo adaptaron. Los pasos son los siguientes:
    - Analizar el formato digital y la documentación del recurso. La salida de este paso es el catálogo de datos y restricciones y la lista de características del tesauro.
    - Definir correspondencias entre los datos de entrada y salida de SKOS RDF. La salida de este paso son tablas de correspondencias de los datos a los elementos del esquema.
    - Desarrollar un algoritmo para el programa de transformación. La salida de este paso es un programa de conversión.

- Aplicar el método. El método se ha aplicado a tres tesauros: *Integrated Public Sector Vocabulary* (IPSV), *Gemeenschappelijke Thesaurus Audiovisuele Archieven – Common Thesaurus Audiovisual Archives* (GTAA) y *Medical Subject Headings* (MeSH) porque se utilizan en la práctica y representan progresivamente un tesauro complejo.
- Evaluación el método. Los casos de estudios muestran que el método da una orientación adecuada en la identificación de las características comunes de tesauros.

Este método transforma el contenido del recurso en un esquema de una

ontología. La ontología resultante se expresa en SKOS RDF / OWL y el método genera una única ontología.

(Wielinga et al., 2001) presentan un método para transformar el *Art & Architecture Thesaurus* (AAT) tesauro en una ontología RDF (S). El AAT es la más elaborada y normalizada base de conocimiento relativa a las clasificaciones de objetos de arte. AAT se publica mediante interfaz Web y también está disponible en formato XML. Este enfoque transforma el contenido del recurso en el esquema de una ontología. El método consta de los siguientes pasos:

- Convertir la jerarquía completa AAT en una jerarquía de conceptos, donde cada concepto tiene una etiqueta correspondiente con el término principal.
- Añadir una serie de conceptos adicionales, por ejemplo, se añadieron los conceptos que representan un estilo o período de tiempo.

La ontología resultante se implementa en RDF(S) y el método maneja una ontología.

Un método para la transformación de tesauros en ontologías es presentado en (Hyvönen et al., 2008). El método se ha aplicado al tesauro YSA y la ontología resultante ha sido alineada a DOLCE. Los autores señalan que, aunque una transformación sintáctica en SKOS (W3C, 2005a) puede ser útil, no es suficiente desde un punto de vista semántico. También afirman que, a menos que el significado semántico de las relaciones de un tesauro sea hecho más explícito y preciso para que pueda ser interpretado por los ordenadores, la versión SKOS es confusa para el ordenador al igual que el tesauro clásico. Por lo tanto, este método de transformación de tesauro a ontología no es sintáctico, se lleva a cabo refinando y enriqueciendo las estructuras semánticas de un tesauro. Este método transforma el contenido del recurso en un esquema de una ontología. Los autores no proporcionan información alguna sobre la implementación del tesauro YSA ni si el proceso es semi-automático o no. El método se basa en los refinamientos semánticos y en las extensiones de la estructura del tesauro:

- Falta de enlaces en la jerarquía *Subclase de*. Las relaciones “Término Genérico”, por lo general, no estructuran los términos en una verdadera jerarquía. Por este motivo, el principio central de este método es evitar la herencia múltiple.
- La ambigüedad de las relaciones de “Término Genérico”. Estas relaciones pueden significar relaciones *Subclase de*, relación *parte de* o relación *instancia de.*

- No transitividad de la relación “Término Genérico”. La transitividad en las cadenas de este tipo de relaciones no está garantizada desde el punto de vista instancia-clase-relación.

La ontología resultante, basada en el tesauro YSA, es la ontología general Finlandesa YSO, que se expresa en RDF(S)/OWL. Por último, la ontología resultante se alinea a la ontología DOLCE.

6. CONCLUSIONES

Los tesauros tienen una importancia estratégica como instrumentos para la organización y gestión de la información, debido a su eficiencia como herramienta para controlar la ausencia de la precisión y ambigüedades propias del lenguaje natural.

A pesar de los beneficios que ofrecen estas formas de organización del conocimiento, las ventajas producidas por el contexto de la Web Semántica junto con los problemas asociados a los tesauros tradicionales reflejan la necesidad del cambio. El cambio parece estar orientado a procesos de reingeniería y/o migración de los tesauros a ontologías a través de RDF y OWL.

Ante esta situación, no debería olvidarse que los tesauros pueden ser considerados una base de conocimiento excelente que puede ser utilizada en el proceso de construcción de ontologías. Por tanto, un tesauro puede tener una estructura jerárquica apropiada, una adecuada base léxica y ser conforme a los acuerdos de los estándares de la Web Semántica (García, 2004).

Desde la perspectiva del proceso de reingeniería, las diversas especificaciones que propone SKOS permite el desarrollo de una infraestructura que puede ser utilizada para interpretar e intercambiar datos de tesauros en el entorno de la Web Semántica.

Desde la óptica del proceso de migración de los tesauros existentes a ontologías, muchos de los actuales problemas de la organización y gestión de la información están siendo solucionados. Asimismo, las ontologías creadas tienen un valor añadido sobre los tradicionales tesauros, consecuencia de la implementación de una semántica más elaborada.

REFERENCIAS

Berners-Lee, T. (1998). Semantic Web Road Map. Disponible en: http://www.w3.org/DesignIssues/Semantic.html

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Disponible en: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Corcho, O., Fernández-López, F., Gómez-Pérez A. (2003). Methodologies, tools, and languages for buildings ontologies. Where is their meeting point? Data & Knowledge Engineering, vol. 46, nº 1, p. 41-64.

Cross, P., Brickley, D., Koch, T. (2000). Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema.

Food and Agriculture Organization of the United Nations (FAO) (1997). AGROVOC. Multilingual Agricultural Thesaurus. AGROVOC flyer. Disponible en: ftp://ftp.fao.org/gi/gil/gilws/aims/references/flyers/agrovoc_en.pdf

García, J. A. (2004) Instrumentos de representación del conocimiento: tesauros versus ontologías. Anales de documentación, nº 7, pp. 79-95.

Gómez-Pérez, A., Fernández-López, M., Corcho, O. (2003). Ontological Engineering. Springer Verlag. p. 403. ISBN: 1-8523355-13.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2), pag.199-220.

Hahn, V. (2003) Turning informal thesauri into formal ontologies: a feasibility study on biomedical knowledge re-use. Comparative and Functional Genomics, 4:94–97(4).

Hahn, U., Schulz, S. (2003). Towards a broad-coverage biomedical ontology based on description logics. pac symp biocomput. pages 577–588.

Hall, M. (2001). CALL Thesaurus Ontology in DAML. Dynamics Research Corporation.

Hyvönen, E., Viljanen, K., Tuominen, J., Seppälä, K. (2008). Building a national semantic Web ontology and ontology service infrastructure -the finnonto approach. In ESWC, pages 95–109.

International Standards Organization (ISO) 2788:1986 (1986). Guidelines for the establishment and development of monolingual thesauri.

International Standards Organization (ISO) 5964:1985 (1985). Documentation - Guidelines for the establishment and development of multilingual thesauri.

Lassila, O., McGuinness, D. (2001). The Role of Frame-Based Representation on the Semantic Web. Technical Report KSL-01-02. Knowledge Systems Laboratory. Stanford University. Stanford, California.

Lauser, B., Sini, M. (2006). From agrovoc to the agricultural ontology service/concept server: an owl model for creating ontologies in the agricultural domain. In DCMI ’06: Proceedings of the 2006 international conference on Dublin Core and Metadata Applications, pages 76–88. Dublin Core Metadata Initiative.

National Information Standards Organization (ANSI/NISO) Z39.19:1993 (1993). Guidelines for the construction, format, and management of monolingual controlled vocabularies.

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S. (2004). Reengineering thesauri fornew applications: The agrovoc example. J. Digit. Inf., 4(4).

Studer, R., Benjamins VR., Fensel, D. (1998) “Knowledge Engineering: Principles and Methods.” , IEEE Transactions on Data and Knowledge Engineering, Vol. 25 (1-2), pp. 161-197.

van Assem, M., Malaisé, V., Miles, A., Schreiber, G. (2006). A method to convert thesauri to skos. pages 95–109.

van Assem, M., Menken, M. R., Schreiber, G., Wielemaker, J., Wielinga, B. J. (2004). A Method for Converting Thesauri to RDF/OWL, presented at ISWC'04, Springer. Hiroshima, Japan.

Vilches-Blázquez, L.M., Martins, B., Wyttenbach, A., Bernabé, M.A., Álvarez, M., Luzio, J., Borbinha, J. (2007). Geographical and historical thesauri: The state of the art. DIGMAP Deliverable D2.1. ECP-2005-CULT-038042, DIGMAP.

Vilches Blázquez, L.M., Rodríguez Pascual A.F., Bernabé Poveda, M.A. (2006a). Ingeniería ontológica: El camino hacia la mejora del acceso a la información geográfica en el entorno Web. Avances en las Infraestructuras de Datos Espaciales. Eds. Granell, C. & Gould, M., Col•lecció Treballs d´Informàtica i Tecnología. Núm. 26. Universitat Jaume I. pág 95 –103. Castellón de la Plana. ISBN: 84-8021-590-9.

W3C (2008a). SKOS Simple Knowledge Organization System Primer. W3C Working Draft. Disponible en: http://www.w3.org/TR/skos-primer/

W3C (2008b). SKOS Simple Knowledge Organization System Reference. W3C Working Draft. Disponible en. http://www.w3.org/TR/skos-reference/

W3C (2005a). SKOS Core Vocabulary Specification. W3C Working Draft 2. Disponible en: http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102

W3C (2005b). SKOS Core Guide. W3C Working Draft 2. Disponible en: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102

W3C (2005c). Quick Guide to Publishing a Thesaurus on the Semantic Web. W3C Working Draft 17. Disponible en: http://www.w3.org/TR/2005/WD-swbp-thesauruspubguide-20050517

W3C (2004). RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation. Disponible en: http://www.w3.org/TR/rdf-schema/

Wielinga, B. J., Schreiber, A. Th., Wielemaker, J., Sandberg, J. A. C. (2001). From thesaurus to ontology. International Conference on Knowledge Capture. Proceedings of the 1st international conference on Knowledge capture. Victoria, British Columbia, Canada. Pages: 194 – 201. ISBN:1-58113-380-4

Wilcox , K., Agre, P. (2005). Semantic Web Technologies in Cultural History: The State of the Art. The CIDOC Conceptual Reference Model.

Wilson, M. (2002). Migrating from Thesauri to Ontologies. Disponible en: http://www.w3c.rl.ac.uk/ukofficepasttalksindex.html

Wilson, M., Matthews, B. (2002). Migrating Thesauri to the Semantic Web. http://www.w3c.rl.ac.uk/pasttalks/slidemaker/XML_UK_SW_Thes/Overview.html