Revista: *Anuario ThinkEPI. Análisis de tendencias en información y documentación.*
Volumen: 7 Páginas: 173-177. Fecha: 2013.

Hacia la primacía de los conceptos sobre los términos en los vocabularios para la web semántica

Por José-Antonio Moreiro
Universidad Carlos III de Madrid
jamore@bib.uc3m.es

Title: *Towards the prevalence of the concepts on the terms in the vocabularies for the Semantic Web.*

Palabras clave: Términos; Conceptos; Palabras-clave; Representación semántica; KOS; Interoperabilidad; Relaciones semánticas.

Keywords: Terms; Concepts; Key words; Semantic representation; KOS; Interoperability; Semantic relations

Cómo eran (y aún continúan siendo) las cosas.

Que términos y conceptos son inseparables en la comunicación se volvió un axioma desde que Sausurre situara el dualismo asociativo de los signos entre la estructura externa y la estructura abstracta. Esta cualidad también aparece en recuperación de la información, aunque de manera más graduada por la profundidad y exactitud que se quiera obtener del léxico. De manera que podemos actuar desde aquellos sistemas que se centran en la extracción o asignación de palabras sin normalizar tomadas del lenguaje natural, hasta los que buscan realizar intercambios más semánticos a partir de los conceptos de los que hablan los documentos, pasando por los que se manejan en torno a los términos de un dominio cognitivo.

No podemos olvidar que el léxico existe porque nos deja intercambiar conceptos mentales. Su fin es semántico, como sucede con la Web cuando busca que el contenido de los documentos sea interpretable por las máquinas para ocasionar significados bien concretos desde discursos muy extensos. Si se quiere alcanzar el significado de los datos hay que moverse en la dimensión abstracta de los conceptos.

Mientras que, también a la hora de representar y organizar la información en Internet, la Web social se basa en procesos más pegados a la palabra, a la estructura externa. Tanto para indizar como para recuperar, el empleo del lenguaje libre o común

se hace sin vocabularios intermediarios. Se busca en texto completo o en indizaciones etiquetadas colaborativamente y se recuperan los documentos que contienen o a los que se les han asignado las mismas palabras de búsqueda utilizadas. Claro que nunca se superan los inconvenientes planteados por la sinonimia y la polisemia, o por la libre combinación de palabras en expresiones sintagmáticas. La utilización de estos significantes es fácil, pero la recuperación resulta dispersa, incoherente y desvinculada. De forma que cualquier búsqueda se puede hacer crónica hasta dar con la expresión cierta, lo que supone un resultado muy incierto y, seguramente, parcial. El empleo sin límite de todas las posibilidades léxicas del lenguaje natural se propagó hace sesenta años para realizar búsquedas en bases de datos y se ha continuado hasta hoy como demuestra la notable presencia de palabras-clave etiquetadas en las folksonomías (Hassan, 2006).

Queriendo evitar las imprecisiones del lenguaje libre, aquella primera oleada informática respaldó asimismo la creación de vocabularios controlados para intercambiar información entre las bases de datos y los usuarios, y lo hizo sobre bases terminológicas (Felber y Pitch, 1984). Así lo asegura la norma UNE 50-106-90 en su introducción cuando dice que se busca representar conceptos mediante términos normalizados y sus relaciones (AENOR, 1990). El origen y el resultado de las representaciones es conceptual, pero se maneja un subconjunto controlado del lenguaje natural, un dominio cognitivo, cuyos significados alcanzan estabilidad al referirse cada término, y solo uno, a un concepto. Esos términos son descriptores sustantivos en los que se nominaliza cualquier otra categoría gramatical.

Los tesauros ordenan cada descriptor dentro de una clase. Muestran así su rigor taxonómico heredado de los esquemas jerárquicos positivistas, si bien facilitan que se pueda indizar y recuperar homogéneamente, por lo que se convirtieron en paradigma de los vocabularios controlados. A la hora de emplear los términos tienen una clara función reguladora. Y, aunque fomentaron su uso combinado y una buena navegabilidad por asociación, mostraban muchos inconvenientes para compartir información en la Web (hasta que aparecieron los SKOS) y para fijar con claridad y exactitud las relaciones.

Cómo han cambiado.

Los tesauros contienen términos inseparablemente de los conceptos que subyacen a ellos (Dextre y Lei, 2012). Si bien los emplean con demasiada rigidez en la descripción, poca exactitud al fijar las relaciones y escaso ajuste con las posibilidades de

relación entre los términos. Cuando se detectaron estas necesidades empezó un progresivo avance hacia la amplitud conceptual que en pocos años se hizo norma, con profundas consecuencias para los nuevos vocabularios.

En algunos aspectos las nuevas normas (ANSI/NISO. Z39.19-2005; BSI Group 2005-2007; e ISO 25964-1-2011) mantienen el carácter terminológico de los vocabularios controlados, en otras lo superan con creces. Lo que menos cambia es la estructuración taxonómica de los términos. Pero ciertamente aquellos ofrecen progresión semántica incluso por la vía expresiva. Es inequívoca la aportación de los anillos de sinónimos para superar la ambigüedad en las búsquedas, con beneficio inmediato al aplicarse a la web social. Establecen equivalencias de las palabras empleadas con sus sinónimos en la fase de indización, delimitando el campo de significación del concepto buscado al emplear todas sus posibles denominaciones. De forma que un concepto se representa por cualquiera de las palabras con significado equivalente. Como la unión hace la fuerza, la recuperación puede hacerse por todos y cada uno de los términos del anillo, sin que ninguno sea preferente. Aunque se vale de los significantes, la suma de todos los posibles que se refieren al mismo concepto facilita la obtención de este, sin necesidad de que el contenido haya sido indizado o esté controlado.

Llegados a este punto, hay que mencionar la base de datos de referencia léxica WordNet, cuya noción básica, el Symset, representa un concepto que puede ser lexicalizado en el lenguaje mediante diferentes palabras. De manera, que Symset viene a significar concepto, pues este se alcanza por la unión de los sinónimos que soportan la relación de equivalencia. Precisamente, a partir de los symsets se establecen en WordNet las relaciones, no entre los lexemas o sentidos individuales.

Conceptos y redes semánticas

Sin embargo, en la progresión hacia el manejo de los conceptos ha tenido mayor trascendencia el establecimiento de las redes semánticas conocidas como Tesauros Conceptuales que se constituyen por conceptos y relaciones entre conceptos. La causa de esta transformación se sitúa en la intención de visualizar los índices para adaptarlos al empleo generalizado de las pantallas. Más allá de alcanzar los conceptos hay que

considerarlos como una propuesta de navegación y de visualización mediante grafos explícitos de conexiones informativas. Por más que en las redes semánticas cada nodo representa un concepto con el que se asocian otros en una malla tejida según las típicas relaciones preferenciales, jerárquicas o asociativas. Las relaciones semánticas que se dan en un texto se simbolizan en ellas mediante grafos compuestos por los conceptos o nodos y por los arcos de interconexión entre los conceptos. De esta suerte, aunque las estructuras taxonómicas fueron los modelos iniciales, las asociativas muestran ahora un notorio auge, cada vez más cercanas a las que se establecen entre los términos del lenguaje natural.

El desarrollo de los tesauros conceptuales en cuanto redes semánticas ha tenido consecuencias inmediatas sobre el léxico y sus relaciones. Al representar en una red semántica los vínculos de significado que se establecen en el texto, y que hasta hace poco se ignoraban en los sistemas de indización, se ha dado un paso más en la progresión de los términos a los conceptos. Está causado por la posibilidad de manifestar la correlación existente entre dos conceptos, mostrando incluso cómo se unen los nodos conceptuales y la dirección de esa relación, en una configuración que debe mucho al modelo de Entidad/Relación. Mediante la sindicación entre nodos, las redes semánticas han prolongado el modo anticipado por los mapas conceptuales en educación.

Activada precisamente por esos enlaces, la identificación verbal de asociaciones funcionales ha sido una de las novedades más esperanzadoras de los últimos años. La potente irrupción de los documentos multimedia forzó a una adaptación rápida de los vocabularios, en especial para indizar imágenes y documentación software. La nominalización de los descriptores generaba mucha pasividad ante las nuevas necesidades, sobre todo a la hora de describir las acciones, componente sustancial del mensaje en ese tipo de documentos. Precisamente porque los verbos expresan acciones mientras que los sustantivos son más estáticos y no tienen capacidad de representar las condiciones de admisibilidad de las acciones. A este respecto, resulta muy descriptivo si pensamos que comer es verbo y que esta acción no puede ser caracterizada por ningún sustantivo sin alterar profundamente su significación ¿Cómo describir a alguien comiendo con el sustantivo comida? Si elegimos esta opción estaríamos expresando la acción por quien la sufre. De este modo, a la hora de representar la información, se precisa designar mediante verbos los estados de hechos, los procesos o las relaciones

entre entidades, como veremos a continuación, que hasta ahora quedaban fuera de los vocabularios combinatorios.

En las redes semánticas, el conocimiento pasó a representarse por frases simples que se estructuraron en tripletas de correlación Sujeto–Verbo–Predicado, en cuanto asociación de un concepto con otro a través de una relación. Este grafo de representación, donde los sujetos y objetos conforman los nodos y las propiedades o predicados determinan el arco de enlace, justifica la participación verbal en la representación de las relaciones existentes entre los conceptos, uno de los hechos más diferenciales a la hora de pasar desde los KOS tradicionales a los SKOS. El empleo de estas frases rompió el monopolio de los sustantivos para representar los conceptos. Con la utilización de los verbos se abría ilimitadamente el número de asociaciones conceptuales consideradas en los vocabularios, al tiempo que se incardinaba con los planteamientos lógicos exigidos para poder intercambiar la estructura de los conceptos entre la mente y los ordenadores. La consecuencia ha sido que no solo se ha multiplicado el número posible de asociaciones entre términos en un vocabulario, si no que además las relaciones entre nodos han permitido precisar el contexto de su significado con mayor exactitud.

El conocimiento representado por frases estructuradas en las redes semánticas ha tenido un evidente paralelismo en la creación de declaraciones con RDF. Las partes de la sentencia o declaración se llaman precisamente Sujeto, que es el recurso o cosa sobre el que versa la declaración; Predicado, la propiedad o característica del sujeto que se expresa mediante esta declaración; Objeto: valor de la propiedad a la que se refiere el predicado. Y se expresan en un grafo unidireccional, en el que los nodos representan a los sujetos y a los objetos, mientras que el arco de relación lo hace con las propiedades o predicados. De esta forma el procesamiento lógico sobre el que se basan los esquemas para la elaboración de ontologías sigue muy de cerca las estructuras gramaticales y la progresión de los vocabularios. No es de extrañar esta coincidencia si la aspiración es alcanzar a procesar el significado mediante la lógica gramatical.

Representación del conocimiento	*Frases estructuradas en grafos*
Redes semánticas	concepto → relación → concepto nodo agente → arco → nodo objeto sujeto → verbo → predicado

Declaraciones con RDF	

Representación del conocimiento por frases estructuradas.

Para mejorar la precisión de las recuperaciones documentales, el funcionamiento de los nuevos vocabularios conceptuales se basa en el diseño de ontologías. Desde ellas se generan, pues permiten distinguir los sinónimos, suprimir los homónimos e inducir relaciones asociativas entre los términos que los integran. En esos vocabularios, los enlaces en la red se establecen tras adaptarse al espacio conceptual hipertextual mediante el lenguaje XML. Se obtiene un corpus terminológico cuya representación se establece como una red semántica neuronal: cada nodo es un concepto semántico con el que se asocian una serie de términos. Ante la pregunta de un usuario se confrontan los conceptos que busca con los elementos de la red terminológica diseñada como mapa representativo de los textos, por lo que actúan como vocabularios que organizan la información de cualquier objeto disponible en la red.

Para representar la información, y frente a lo común en los vocabularios tradicionales, las ontologías no cuentan con una organización acotada y precoordinada de relaciones semánticas. Por lo que necesitan un mayor nivel de descripción del vocabulario que los tesauros y también mayor desarrollo semántico de las relaciones entre los conceptos. Pues permiten que cualquier relación semántica clarificada se incorpore a su estructura, de acuerdo con la necesidad de representación marcada por la propia dinámica del conocimiento.

¿Hacia dónde van los VES?

Es cierto que el conocimiento continúa organizándose en taxonomías, como demuestra su presencia en los esquemas de clasificación, en los tesauros, en los modelos conceptuales y en las ontologías**.** Sin embargo los vocabularios conceptuales y ontológicos están muy lejos de limitarse a cualquier listado taxonómico. La preocupación es dotar de significado a los objetos de información para que puedan ser entendidos por las máquinas. Por eso hay que trasladarles los términos con exactitud,

hay que facilitarles los enlaces entre vocabularios existentes y hay que hacerles comprender todas las relaciones existentes entre sus elementos. En definitiva, se va hacia vocabularios más interoperables, que especifican cómo un concepto se relaciona con otros conceptos.

A la hora de representar el conocimiento es cada vez más habitual valerse de ontologías para relacionar los conceptos pertenecientes a individuos de clase que por su medio se constituyen en bases de conocimiento. La ontología describe explícita y formalmente los conceptos (clases) de un marco de conocimiento compartido, así como las propiedades que especifican sus características y atributos (*slots*), junto con restricciones sobre los slots.

Para hacer posible que distintas aplicaciones Web intercambien su información, RDF ha fomentado la reutilización. A la hora de buscar un término correspondiente entre dos o más ontologías se parte de una fuente y se alcanza una de destino (BSI Group, 2007). A la hora de desarrollar ontologías se pueden aprovechar otras organizaciones conceptuales preexistentes, como tesauros, de lo que se ocupan metodologías como Methontology. Este proceso se realiza sin pérdida alguna de significado, de manera que el conocimiento inicialmente dispuesto para un contexto puede aplicarse en otro distinto. Los sistemas tienen que poder intercambiar automáticamente información, pues todas las propiedades tienen un significado bien establecido (Interoperabilidad semántica) con una sintaxis XML para los documentos más habituales (interoperabilidad sintáctica). De manera que cualquier aplicación podrá usar un concepto relacionado con los que maneja aunque no estuviese incluido en su ontología.

Es incontestable acudir a la interoperabilidad para elaborar o utilizar vocabularios, tanto para acceder a la información precisada como para representar el contenido específico de un campo. El encargo de interoperabilidad entre KOS se origina tanto en la tecnología de la Web que favorece el intercambio de información en tamaño hasta hace poco ni barruntado, como en la contingencia de reutilización y lucro de los recursos informativos que primero se hicieron para un servicio y uso concretos. Esta forma de actuación se determina desde la propia aspiración de globalidad que tienen los KOS, junto a la debida accesibilidad social de la información en el sector público, así como desde la existencia de repertorios indizados mediante algún KOS y especialmente con descripciones de metadatos. De manera más matizada, le incumbe a la

interoperabilidad hacer eficaces los términos de un recurso en otro, o tomar vocabulario de un recurso para utilizarlo con éxito en otro. En especial, busca conjuntar varios recursos en otro mayor de cara a lograr unos mapas conceptuales y terminológicos más comprensibles (Méndez y Greenberg, 2012).

El hecho de que muchos vocabularios, taxonomías y ontologías estén en formato electrónico facilita su importación a otro sistema. SKOS ofrece un esquema para codificar vocabularios controlados en XML y migrarlos al entorno de la web semántica. Esto no es solo útil para publicar vocabularios, sino como mecanismo para representar relaciones entre distintos esquemas conceptuales. Disfrutando así de la capacidad de relacionarlos a partir de su similitud semántica y de las relaciones reales que existen entre las entidades que los crearon y gestionan. En estas condiciones, reutilizar es un modo de compartir información para mejorar su representación en la web.

No habrá Web semántica sin recuperación semántica, sin poder efectuar la búsqueda por conceptos, por significados, por ideas. Por ello el software tiene que distinguir entre ideas y términos. La captación de los conceptos se hace desde la lógica formal, pero también desde el propio léxico cuyas posibilidades de tratamiento han aumentado notablemente en las últimas normas internacionales.

Referencias

AENOR (1990) *Directrices para el establecimiento y desarrollo de tesauros monolingües*. UNE 50-106-90. Madrid: AENOR, 1990.

ANSI/NISO. Z39.19-2005. *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. Bethesda, Maryland: NISO Press, 2005.

BSI Group (2005-2007). *Structured vocabularies for information retrieval: guide*. London: BSI (BS 8723/1-4); ISO 25964-1 (2011). Information and documentation - *Thesauri and interoperability with other vocabularies* - Part 1: Thesauri for Information Retrieval. ISO 25964-1:2011.

Dextre Clarke, Stella G., Lei Zeng, Marcia (2012). From ISO 2788 to ISO 25964: the evolution of Thesaurus standars towards Interoperability and Data modeling. *Information Standards Quarterly*, v. 24, nº 1. pp. 20-26.

Felber, H. y Pitch, H (1984). *Hispanoterm. Métodos de terminografía y principios de investigación terminológica*. Madrid: CSIC-Instituto Miguel de Cervantes.

Hassan Montero, Yusef (2006). Indización Social y Recuperación de Información. En: *No Solo Usabilidad*, nº 5. <nosolousabilidad.com>.

ISO 25964-1 (2011). Information and documentation - *Thesauri and interoperability with other vocabularies* - Part 1: Thesauri for Information Retrieval. ISO 25964-1:2011.

Méndez, Eva; Greenberg, Jane. "Linked data for open vocabularies and HIVE's global framework". *El profesional de la información*, 2012, mayo-junio, v. 21, n. 3, pp. 236-244.