

PRESENTACIÓN JORNADAS SOBRE LENGUAJES Y GESTIÓN DE INFORMACIÓN

Vocabularios controlados e indización conceptual

INSTITUTO CERVANTES 17 DE JUNIO DE 2010

Agradezco a los organizadores la invitación para tomar parte en esta primera Jornada Profesional sobre Lenguajes y Gestión de información.

Introducción

La primera parte de este Encuentro se va a dedicar, acertadamente, a los vocabularios controlados y a la indización conceptual, lo que va a permitir trazar una panorámica global y coherente sobre la evolución de los lenguajes y de los sistemas de organización del conocimiento, desde los primeros tesauros de mediados del siglo XX, hasta nuestros días, donde modelos con metodologías tomadas de los campos de la Informática y de la Inteligencia Artificial han generado otro tipo de sistemas que esencialmente no difieren mucho de los anteriores aunque formal y metodológicamente presentan grandes cambios. Todos ellos están profundamente implicados con el lenguaje y con todos los problemas que su tratamiento documental conlleva y todos ellos están comprometidos de forma especial con el significado, en estrecha relación con la relevancia de los contenidos documentales independientemente de los métodos que se utilicen para garantizarlo, algo ampliamente conocido por los métodos considerados como más tradicionales y que últimamente es una búsqueda incesante de los nuevos sistemas, los empeños puestos en la Web semántica es un claro exponente. Todos tienen la finalidad de gestionar de la información lo más eficientemente posible y una de las claves de su éxito es, en mi opinión, el control del significado, velar por la univocidad de la terminología, y, en definitiva, la profundización en la semántica de dichos sistemas.

Espero que todo lo que se diga aquí esta mañana permita poder establecer esa línea conceptual continua que existe en esta evolución de los lenguajes y de los sistemas de gestión de información, aunque a veces no se sea consciente de ello. No hay, en mi opinión, un divorcio entre unos y otros sistemas en lo esencial sino que asistimos a algo tan natural como el impacto del desarrollo tecnológico, tan influyente y tan condicionante en nuestra especialidad (Hjorland, 2003) y que se ha asumido ese reto

que ha dado lugar no sólo a una generación de sistemas que responden a este nuevo contexto, sin duda deudor de aportaciones previas, sino a actualizar, a introducir el componente tecnológico en los sistemas más tradicionales

Los Lenguajes Controlados

La mayor parte de la terminología que usamos cuando nos referimos a los llamados lenguajes documentales es de reciente cuño, segunda mitad del siglo veinte, y por tanto posterior a la publicación de teorías sobre su construcción, cuyos primeros intentos formales se originan a finales del siglo XIX hasta la mitad del pasado siglo más o menos con alguna aportación posterior. Cuando nos referimos a los lenguajes controlados, enunciamos sus características y sus métodos, lo hacemos a partir de una reflexión a posteriori sobre la producción de lenguajes ya existentes que fueron generados de forma especulativa, es decir, al margen de modelos teóricos globales que se conforman a partir de los Proyectos Cranfield de finales de los años cincuenta del siglo pasado (Ellis, 1992). Este hecho no resta importancia a las contribuciones teóricas de esta época que dan sentido y propician el estado actual en buena medida, basta recordar las aportaciones de Ranganathan con su teoría de las facetas que es utilizada ampliamente por toda clase de sistemas en la actualidad, pero no poseen la coherencia, la capacidad de reflexión y la garantía que otorga investigar bajo un determinado prisma teórico y metodológico.

Se hizo, pues, un gran esfuerzo sintetizador a partir de los años sesenta de las aportaciones hechas previamente y surgen así los primeros estudios globales de los lenguajes documentales de marcado carácter lingüístico, Hutchins es un buen ejemplo, y también su clasificación bajo distintos criterios, uno de los cuales fue el control que los dividía en lenguajes controlados y libres. Creo que es necesario reflexionar sobre esta clasificación. Un lenguaje controlado es un lenguaje normalizado formal y semánticamente, independientemente del método que se utilice para ello. Se puede controlar el significado de una expresión de forma manual o por medio de otros parámetros como pueden ser los algoritmos pero el objetivo es controlar el lenguaje, controlar el significado porque la excelencia del sistema tiene mucho que ver con este control. Hay una tendencia a identificar los lenguajes controlados sólo con los sistemas más tradicionales pero hoy no podríamos decir lo mismo, por más que en su inicio así fuera. Habría que reflexionar sobre el concepto de control a la luz de nuestros días

Asimismo, la elección del término libre como opuesto a controlado tampoco me parece muy afortunada porque de las explicaciones que se dan de este tipo de lenguajes se deduce que siempre se habla de un listado cerrado de términos que representan un dominio temático determinado, lo que significa que no todas las expresiones son válidas para representarlo y que el sólo hecho de tener una lista cerrada significa control aunque de menor grado. Esta situación se genera porque trabajamos con un metalenguaje que es representativo del primer nivel lingüístico y que se interesa sólo por la relevancia del documento primario y, por lo tanto, tiene que ser selectivo de los conceptos y de la terminología que incorpora. Se puede este hecho trasladar a sistemas menos tradicionales que parten del documento a texto completo, que incluso pueden no tener lista alguna previa, pero que, sin duda, necesitan de mecanismos para efectuar ese control porque el objetivo de la univocidad de los lenguajes que se utilizan en los sistemas de información es de alguna manera compartido ya que está en estrecha relación con su eficacia en la recuperación de la información. Los métodos para conseguirlo varían pero los objetivos permanecen. Por tanto, es necesario que revisemos todas estas cuestiones a la luz de lo que pasa en la actualidad y no nos conformemos con clichés obsoletos que tal vez tuvieron sentido en una época.

La indización conceptual

La identificación de conceptos para la elaboración de vocabularios controlados comienza con el análisis de documentos especializados a partir de los cuales se extraerán aquellos que sean relevantes para representar su contenido. Un punto clave en esta etapa es el concepto de relevancia con que nos aproximemos al documento porque todo lo demás será inexistente para el sistema. Aparte de pautas referidas a la exhaustividad y otras derivadas de las necesidades específicas del sistema de información concreto al que va a servir el lenguaje, sería conveniente plantearse también cuestiones referidas al diseño del sistema y del lenguaje que se va a construir, porque va a afectar directamente a la indización y porque está en directa relación con los modelos teóricos generales de la Documentación. Es verdad que toda obra se asienta sobre teorías aunque no se hagan explícitas y aunque, incluso, quien escribe no sea muy consciente de ellas. Lo que quiero poner manifiesto aquí es la necesidad de una reflexión consciente sobre el diseño, previa a la construcción del sistema, que debería ser parte del proyecto. Dependiendo de dónde nos situemos, los elementos de diseño serán unos u otros y esto afectará directamente a la indización. Por ejemplo, si

indizamos conceptos representativos de los temas tratados en un texto, estaríamos acogiéndonos a modelo objetivo o físico (es el habitual), pero no se pueden olvidar las otras propuestas procedentes de teorías cognitivas y sociocognitivas para las que la relevancia no sólo es objetiva. Estas últimas amplían el concepto de relevancia, lo que incide directamente en la indización, al tiempo que consideran al usuario como un elemento activo de diseño, lo que también afectaría a la indización. Luego sería aconsejable reflexionar sobre estos factores antes de iniciar el trabajo en sí. La mayor parte de lo que se publica se centra en el proceso y los métodos, pero poco o nada se dice del diseño y de sus consecuencias en el producto final y habría que considerarlo

Centrándonos en el proceso de la indización y de la elaboración de los lenguajes para la gestión de la información, pueden distinguirse dos metodologías claramente diferenciadas: la aproximación terminológica en el proceso de la indización que conlleva el uso de métodos cuantitativos donde la unidad de análisis primaria es la grafía del signo lingüístico y se identifica con métodos automáticos y la aproximación conceptual donde la unidad de análisis es el concepto, implica el empleo de métodos cualitativos o mixtos y se identifica con métodos manuales o semiautomáticos. Así vemos que, mientras en el primer caso la tarea primordial es identificar, a partir del término o expresión gráfica de un corpus documental, el significado aplicando para ello todos los mecanismos automáticos de control necesarios para lograr la univocidad dentro de lo posible, en el segundo caso se persigue la identificación de conceptos relevantes de los contenidos analizados que después se expresarán con la grafía más adecuada. De manera que se ve claramente que son caminos invertidos los transitados por uno y otro método aunque buscando, sin duda, un control semántico. Vamos a centrarnos brevemente en la indización conceptual que nos ocupa ahora.

El método conceptual tiene amplias implicaciones. El concepto es no sólo la unidad del vocabulario sino también la unidad de análisis para la construcción de la estructura que todo sistema de recuperación debe tener sobre la que se construirá el sistema relacional. Si queremos crear un sistema de información y no sólo vocabularios no estructurados, deberíamos ir más allá y plantearnos que detrás de un concepto hay más que su significado. Los conceptos, en este contexto, no se pueden entender sólo como un continuo, su significado, sino también como un conjunto de características exclusivas, la suma de las cuales hacen que ese concepto sea lo que es y no otra cosa. Esta

concepción se basa en la idea manifestada por Dahlberg de que “cada definición o frase cierta sobre un cierto ítem de referencia genera un elemento de conocimiento acerca de él junto con una característica de su concepto. La suma de las frases necesarias sobre ese ítem de referencia forma la totalidad de las características de su concepto, presenta de forma distintiva su contenido” (Dahlberg, 2009, 171), posteriormente es necesaria una designación que puede ser un término, un código, etc. Es éste un modelo que está en estrecha relación con la denominada aproximación onomasiológica en la representación de los conceptos que ve el concepto no sólo con un contenido semántico, su significado, sino también como la suma de las características antes enunciadas. Esto hace del concepto una figura polivalente activa en los dos momentos más importantes del proceso: la indización y la construcción de la estructura y su corpus relacional.

Hay trabajos que demuestran este método llegando con él hasta el diseño real de una estructura que resultó tener características muy ventajosas como capacidad de polirrepresentación, elevada precisión y exhaustividad y la posibilidad de numerosas relaciones conceptuales porque la misma estructura ponía de manifiesto relaciones que aparentemente estaban ocultas y por tanto indetectables (López-Huertas, 1997)

A continuación vamos a tener la oportunidad de conocer casos concretos de elaboración de este tipo de sistemas que ilustrarán con detalle todo el procedimiento, con el interés de responder a ámbitos bastante distintos del conocimiento. En esta breve presentación sólo he querido hacer énfasis en algunos puntos del tema que nos ocupa que me han parecido de mayor interés.

DAHLBERG, Ingetraut (2009). Concepts and terms- ISKO's major Challenger. *Knowledge Organization*, 35 (2/3), pp. 169-177

ELLIS, David (1992). The physical and cognitive paradigms in information retrieval research, *Journal of Documentation*, v.48 (1) p.45-64

HJØRLAND, BIRGER (2003). Fundamentals of Knowledge Organization. *Knowledge Organization*, 30 (2): 87-111

LÓPEZ-HUERTAS, María José (1997). Thesaurus structure design: a conceptual approach for improved interaction». *Journal of Documentation* 53 (2), pp. 139-177