INTEGRACIÓN DE TEORÍAS PARA LA REPRESENTACIÓN Y RECUPERACIÓN DEL CONOCIMIENTO.

Autores: Dr. Miguel-A. López Alonso
Facultad de Biblioteconomía y Documentación,
malopalo@alcazaba.unex.es

Resumen: Se estudia la larga evolución necesaria para la disolución de las diferencias básicas entre los diferentes Lenguajes Documentales y su relación con la Terminología. Se parte de las Gramáticas Textuales aplicadas al Análisis del Texto y de la adopción del Modelo Cognitivo en los Sistemas de Procesamiento de la Información. Además, se comparan las estructuras conceptuales de representación del conocimiento más utilizadas: clasificaciones, tesauros y ontologías. Se propone la interconexión de los Tesauros Conceptuales y las Ontologías almacenados en Internet, para la mejora de las recuperaciones en la Web.

Palabras clave: Lenguaje Natural.- Lenguajes Controlados.- Sistemas de Clasificaciones.- Tesauros Conceptuales.- Ontologías.- Modelo Conceptual.-

Abstract: It is studied the long necessary evolution for the dissolution of the basic differences between the different Documentary Languages and its relationship with the Terminology. From the Textual Grammars applied to the Textual Analysis and the adoption of the Cognitive Model at the Information Processing Systems. Moreover, the more used knowledge conceptual representation structures: classifications, thesauri and ontologies are compared. It is proposed the interconnection of the Conceptual Thesauri and Ontologies stored in Internet, in order to improve the retrieval on the Web.

Key words: Natural Language. Controlled Languages. Classifications Systems. Conceptual Thesauri. Ontologies. Cognitive Model.

Introducción

En el ámbito de la Documentación Científica, se percibe una tendencia doctrinal al abandono de la visión anatómica de la información como “soporte u objeto físico” desde el punto de vista del sistema de información, y a su sustitución por una visión conceptual como “conocimiento o discurso cognitivo” desde un enfoque centrado en las necesidades del usuario final. Pensamos que el siglo XXI depara a los profesionales de las Ciencias de la Información, como "mediadores en el procesamiento de ésta”, la ingente tarea de remover los obstáculos de sus distintas fases cognitivas; especialmente los derivados de *la organización del conocimiento* y *el desarrollo de las herramientas conceptuales para su posterior recuperación* con el lenguaje de los usuarios.

En la primera parte de este trabajo, se revisa la evolución hacia los modelos conceptuales que se desarrollan a partir de las Gramáticas Textuales aplicadas al Análisis del Texto. Se sigue la tesis conceptual-lingüística de los especialistas en la construcción de Lenguajes Controlados que, como Maniez o Dewèze, consideran la "naturaleza simbólica del signo lingüístico" como elemento común entre los lenguajes naturales y los controlados, y a "sus términos" como la representación de la noción semántica o concepto del que hablan los documentos.

A continuación, se analiza el “estado actual” del antagonismo clásico entre el Lenguaje Natural y los Lenguajes Controlados, como medio de redacción de las ecuaciones de búsqueda en los sistemas de recuperación de la información. Se sostiene la unión de los puntos fuertes de ambos tipos de lenguajes para contrarrestar sus puntos débiles, optando por “una complementariedad” que mejore la precisión de las recuperaciones, haciendo depender sus fortalezas y sus debilidades del contexto en que el usuario busca la información; y se mencionan los resultados de algunos de los más recientes experimentos con dichas herramientas terminológicas.

Finalmente, se estudian las estructuras de representación de la información más utilizadas: clasificaciones, tesauros, ontologías, etc. Se defiende “su convergencia” y la necesaria reutilización de las clasificaciones como punto de partida para la autogeneración de Tesauros Conceptuales y ontologías especializadas.

Se propone la interconexión de las Ontologías y los Tesauros Conceptuales almacenados en las redes distribuidas de Internet, como redes tautológicas para la mejora de la precisión de las recuperaciones en los interfases de usuario de los Sistemas Integrados de Recuperación de la Información de la Web.

1. EL MODELO CONCEPTUAL-LINGÜÍSTICO EN LA CATEGORIZACIÓN DEL CONOCIMIENTO.

A mediados de este siglo se había llegado a un establecimiento del “modo correcto” de concebir los conceptos, categorías y clasificaciones, como revisión de la posición clásica de Aristóteles. Sin embargo, en las últimas décadas con el desarrollo de la Ciencia Cognitiva las teorías sobre la categorización del mundo han sufrido numerosos ataques, y se puede afirmar

que su concepción clásica ha sido reemplazada por *una concepción natural de sus conceptos* con un enfoque centrado en el individuo.

El problema conceptual es determinar si la capacidad humana permite agrupar, de manera unívoca, elementos como miembros de una categoría y luego distinguirla de otras categorías pertenecientes al mismo dominio general. Como resultado, ha ido imponiéndose que esa categorización es relativamente arbitraria, y que la cultura determina las entidades o clases que el clasificador humano discrimina.

De la síntesis de los postulados de varias disciplinas pertenecientes a la Ciencia Cognitiva, la psicología, la antropología y la filosofía, ha ido surgiendo la solución. La categorización no tiene nada de artificial o inmutable, sino que se basa en la información procedente del mundo natural. Las categorías o clases facetadas poseen una estructura interna de estereotipos o marcas de clase, y los demás miembros de la estructura están más o menos próximos a ella según el grado en que comparten determinados rasgos comunes con el elemento que da nombre a la clase.

Como los esquemas clasificatorios no reflejan las relaciones individuales existentes entre los términos, desde un punto de vista cognitivo, se les ha tratado de asimilar con los tesauros facetados postcoordinados. En éstos, los campos temáticos se subdividen en clases de materias (por agrupación de numerosas subclases de facetas), que se eligen por poseer una característica común divisoria.

Dichos esquemas proporcionan un orden jerárquico a muchos pequeños subconjuntos de conocimientos, que se toman como punto de partida para la ordenación de los descriptores en los tesauros. Se trata de obtener el mismo grado de detalle con el planteamiento de "abajo-arriba" del análisis terminológico postcoordinado (segmentación conceptual) que con la clasificación precoordinada de "arriba-abajo" (taxonomía documental).

1.1. Procesos innatos en las tareas de clasificación e indización:

En un sistema del lenguaje, las bases teóricas de las distinciones dicotómicas entre lengua-habla y eje paradigmático-eje sintagmático fueron desarrolladas por Sausurre[1]. De la articulación de ambos ejes se entiende el proceso de un sistema conceptual o lenguaje para generar sistemas físicos, los textos en Lenguaje Natural; de como a partir de estos textos se genera un sistema conceptual, el sistema de clasificación; y de como este sistema genera un sistema físico o notación clasificatoria[2].

A pesar de la contribución del estructuralismo de Sausurre al estudio del significado del signo lingüístico, será Chomsky con su gramática generativo-transformacional quien mejor encuadre las cuestiones propias de la ciencia cognitiva[3]. Éste nos sugiere que el estudio del lenguaje debe incorporarse a un estudio más general de la psicología humana, que nos lleve a una ciencia cognitiva integrada.

Como superación de la orientación atomista de los elementos del lenguaje de Sausurre, Chomsky irá de lo general a lo particular, dando una gran importancia a la forma en que los sujetos estructuran y generan el lenguaje, a partir de lo que él denomina “mecanismo innato” de *estructuras*

mentales de la especie humana. Dará gran importancia a las propiedades de las palabras dentro de la oración (regidas por reglas), encontrando afinidades de la estructura profunda con su significado.

Para avanzar en las concepciones semánticas del lenguaje, se ha recuperado la Ciencia del Texto de Van Dijk (como una teoría interdisciplinar integrada en la Ciencia Cognitiva) en la que la organización estructural del texto se describe en dos niveles, superficial y profundo, siguiendo las estructuras avanzadas inicialmente por Chomsky.

Van Dijk[4] postula como tesis que, durante el "análisis del texto", tienen lugar dos procesos mentales interrelacionados:

- Uno de importancia normal, según el significado del texto para un potencial usuario (microestructuras del texto), y
- Otro de importancia extracontextual, según los elementos del texto que el clasificador induce necesarios para su comprensión (macroanálisis del texto), que suele venir determinado a su vez por: la finalidad clasificatorio-ideológica del sistema de clasificación utilizado, y por las características del sistema de información.

A partir de estos niveles del documento primario podremos describir dos estructuras del sistema de clasificación: una macroestructura global constituida por un sistema de conocimiento determinado y una microestructura variable según el modelo del sistema de clasificación (ej.: enumerativo, facetado, etc.) y de sus correspondientes códigos notacionales.

Por tanto, durante el análisis documental la clasificación opera con objeto de organizar el conocimiento de acuerdo con un esquema conceptual y unificar las partes del documento, situándose al nivel más sintético posible, mientras que la indización adopta una perspectiva analítica o de disección frente al documento[5].

1.2. La Clasificación Documental desde una perspectiva cognitiva:

Jennifer Rowley[6] define la Clasificación Documental como:

> "la disposición de un conjunto de documentos en grupos diversos pero relacionados entre sí por su contenido temático, a partir de la aplicación de un Sistema de Clasificación previamente elegido...";

al que a su vez define como:

> "una estructura metódica de clases u objetos (paneles, facetas, etc.) que agrupa conceptos ordenados del conocimiento, ligados entre sí por unos caracteres comunes, establecidos según unos principios o reglas prefijadas a las que llamamos Lenguajes Documentales".

Y si, también, afirmamos que los Sistemas de Clasificaciones Conceptuales (al igual que el proceso cognitivo humano de clasificar):

> "no se limitan a identificar y describir la forma y el contenido de los documentos, sino que se dedican sobre todo a organizar el conocimiento que albergan, mediante su clasificación y ordenación...";

y que por ésta última se entiende:

> "la acción de proceder, dentro de cada clase, grupo de clases o de modo global, a la disposición de los documentos en una sucesión,

siguiendo un criterio de relación predefinido, único y uniforme (alfabético, numérico o alfanumérico)..."[7].

Llegamos a la conclusión de que entre el enfoque lingüístico de la Ciencia del Texto de Van Dijk, y el enfoque de la Ciencia de la Información de los documentalistas no existe una diferencia fundamental.

Vemos que lingüistas como Zsilka[8] y documentalistas como Svenonius defienden la existencia de un componente subjetivo en toda Clasificación Documental que conduce a que cada clasificador pueda obtener una clasificación diferente de una misma realidad, igual de correcta y útil, según:

- La adaptación de la clasificación a las necesidades que debe servir (la realidad cultural extralingüística, etc.), y
- La porción del conocimiento general que se quiere organizar.

Esto nos lleva a la visión, de Shreider y Uspensky[9], de los Sistemas de Clasificaciones como un tipo de desarrollo "ad hoc" de Gramática Textual Aplicada, que proporciona la acomodación de los documentos que se quieren clasificar, y los mantiene unidos interconceptualmente.

La teoría clasificatoria conceptual toma las ideas de los Sistemas de Clasificaciones Conceptuales difundidos por los seguidores de Ranganathan[10], y defiende que los catálogos conceptuales proporcionan al usuario una mayor utilidad que los alfabéticos. Sin embargo, la aproximación lingüística se apoya en los estudios de Neumann[11], y utiliza la Organización del Conocimiento para la Recuperación de la Información, a partir del desarrollo previo de corpus terminológicos o léxicos controlados que ayuden al procesamiento del Lenguaje Natural.

Aunque existen diferencias importantes entre ambas aproximaciones, propugnamos imprescindible su fusión de manera que la teoría de la Organización del Conocimiento aporte a la teoría conceptual de la Recuperación de la Información sus potentes estructuras del conocimiento ("gestalt structures")[12].

Lo expuesto en este apartado, nos lleva a abordar el controvertido tema de la complementariedad del Lenguaje Natural y de los Lenguajes Controlados, y en cómo afecta a la indización y a la recuperación de la información documental.

2. BREVE HISTORIA DE UNA CONTROVERSIA.

Desde la publicación de las conclusiones de Cleverdon en sus experimentos de Cranfield (1962-1966), y la expansión de memoria y capacidad de almacenamiento alcanzada recientemente por los ordenadores, se ha pensado en una progresiva sustitución de los Lenguajes Controlados por el Lenguaje Natural. Como consecuencia, estas conclusiones han sido minuciosamente revisadas, y diferentes investigadores han demostrado la complementariedad de dichos lenguajes.

Maniez, en la primera parte de su tesis doctoral, desarrolla la cuestión de ¿cuáles son las diferencias y similitudes entre los Lenguajes Naturales y los Lenguajes Controlados?. Para él, "la naturaleza simbólica del signo lingüístico" es el elemento común. El usuario no busca términos de indización para sí mismos sino para los documentos, donde los términos representan al

concepto. La estructura de los Lenguajes Controlados es una restricción de la de los Lenguajes Naturales[13].

Fugmann argumenta:

> "El éxito de una búsqueda documental depende de la acertada predicción de la forma de expresión con que los autores de los documentos han tratado un determinado tema"[14].

Y nos lleva a creer que, las conclusiones de los primeros estudios de Cranfield que asociaban una relación directa entre una alta recuperación y una baja precisión con las búsquedas en Lenguaje Natural, por oposición a las búsquedas con Lenguajes Controlados, fueron consecuencia del uso de metodologías poco sofisticadas, y propone su repetición para verificar si se producen resultados diferentes con los métodos actuales de medición de la relevancia.

Dubois coincide en que:

> "Cualquier intento de revisión de los factores que operan detrás de los citados resultados, requeriría una revisión de los criterios de análisis de las técnicas metodológicas de las que se afirman ventajas y desventajas: semántica, contextual, relacional, materia, comportamiento humano, etc."[15].

Y propone que, para cada enunciación de preguntas en un nuevo entorno de estudio, se establezcan "a priori" las técnicas que identifiquen sus variables más importantes, independientemente del tipo de lenguaje utilizado.

Por el contrario, según Svenonius:

> "La causa de que de la utilización de un vocabulario controlado en las búsquedas se obtenga mayor precisión que con el Lenguaje Natural, se debe a que el esfuerzo intelectual de asignar descriptores manualmente a los documentos predice mejor su relevancia que la asignación automática de cualquier término extraído de un título o de un resumen"[16].

En un Sistema Integrado de Gestión de la Información, con utilización compartida de Lenguajes Controlado y Natural, los puntos fuertes del primero compensan las debilidades del segundo y viceversa.

Un caso típico de complementariedad lo constituye el método de tratamiento de la documentación en los Centros de Prensa, donde se utiliza un vocabulario sectorial polijerárquico perteneciente a varias disciplinas. En esta línea, Kristensen propone que éste tipo de vocabulario postcontrolado se use en la fase de recuperación de la información de prensa, ya que sus documentos con texto completo no se suelen indizar, debido a su gran tamaño y al volumen generado diariamente[17].

2.1. Complementariedad del Lenguaje Natural y los Lenguajes Controlados.

En la actualidad, los sublenguajes controlados tienden a generarse automáticamente, a partir del procesamiento del Lenguaje Natural de los documentos. El vocabulario puede tener una mínima estructura del tipo de

referencias cruzadas o de tablas correlacionadas, y se pueden incluir tesauros multilingües para su uso en las Bases de Datos Internacionales.

Para Lancaster un vocabulario postcontrolado puede ser construido con un esfuerzo equivalente al de un tesauro convencional. Esta opinión se refuerza con las investigaciones de Dubois, Fugmann o Svenonius, que han demostrado la complementariedad de los lenguajes naturales y controlados.

Se considera con Kristensen y Larsson[18], que las recuperaciones que utilizan vocabularios postcontrolados generados automáticamente a partir del Lenguaje Natural, obtienen muchas de las ventajas de los Lenguajes Controlados tradicionales y evitan la mayoría de los problemas lexicales de su uso directo: sinonimias, homografías, etc.

Diversos experimentos en que los usuarios son apoyados, en la enunciación de sus ecuaciones de búsqueda, con términos adicionales extraídos de un tesauro diseñado específicamente para la recuperación con Lenguaje Natural en grandes Bases de Datos; han aportado avances significativos en el conocimiento intrínseco de la relevancia de las recuperaciones[19], y se ha llegado incluso a doblar la precisión en el número de documentos recuperados si el usuario selecciona y usa los términos sugeridos por un tesauro como adicionales a sus propios términos[20].

Otro análisis reciente revela las razones por las que hay tanta diferencia de registros (entre el 37% y el 48% con Lenguaje Natural, frente al 49% y el 86% con descriptores, respectivamente) entre el uso del Lenguaje Natural de los títulos de los artículos y el lenguaje controlado de los descriptores[21]. La primera es que aunque la mayoría de los títulos tratan de expresar el contenido de los artículos, no siempre aportan las suficientes palabras clave para formular una ecuación de búsqueda completa. La segunda y la más importante es que los títulos no alcanzan a expresar las diferentes formas de representar un tema concreto, necesitando un vocabulario normalizado que amplíe el Lenguaje Natural del usuario.

3. ESTRUCTURAS CONCEPTUALES DE REPRESENTACIÓN DE LA INFORMACIÓN.

Las clasificaciones temáticas representaron un gran adelanto frente al mero orden alfabético de los diccionarios y léxicos tradicionales, pero están limitadas al no poder reflejar las relaciones que existen entre los términos individuales. Sin embargo, los Sistemas de Clasificaciones en uso en los grandes Sistemas Bibliotecarios fueron construidos de acuerdo con los principios de una metodología enumerativa, y aunque después de la II Guerra Mundial hubo un relativo auge en el desarrollo de los sistemas basados en la metodología facetada, no se ha dado el salto esperado a éstos nuevos sistemas, debido principalmente a problemas organizativos de éstas grandes instituciones.

Tanto Ranganathan como Bliss propusieron unas consistentes y bien fundamentadas teorías conceptuales, asociadas con la aplicación de las facetas a los Sistemas de Clasificaciones Bibliotecarios. No obstante, las limitaciones iniciales no se han eliminado totalmente con la utilización de las subclases facetadas; resulta obligado complementarlas con otro lenguaje

controlado de relaciones más complejas, el Tesauro Conceptual, como punto de encuentro entre los Sistemas de Clasificaciones y las Ontologías Conceptuales.

Se puede afirmar que los Sistemas de Clasificaciones Mixtos o Universales se seguirán usando durante muchos años en las bibliotecas con soporte papel, para el acceso directo a las fuentes primarias. Su supervivencia definitiva dependerá de su progresiva incorporación a los sistemas en línea, con un formato universal informatizado[22].

3.1. Los Tesauros Conceptuales como superación de los Sistemas de Clasificaciones.

Para intermediar en la comunicación entre la Ciencia de la Documentación y la Lingüística Documental, aparecerán los Lenguajes Controlados de estructura combinatoria o Lenguajes Controlados que tratarán los documentos fuera del alcance de los Sistemas de Clasificaciones Universales[23] y se especializarán por áreas del conocimiento. La clasificación, como ayuda a la catalogación y a la recuperación por materias, ha sido sustituida por la indización mediante descriptores tomados de los tesauros y la recuperación mediante su combinación con operadores booleanos.

Los Sistemas Comerciales de Recuperación de la Información Documental han recurrido a la utilización de los descriptores de los tesauros de cada área del conocimiento para indizar los documentos por sus títulos o resúmenes y, tras la ponderación de los conceptos más repetidos los han utilizado posteriormente como descriptores postcoordinados en la recuperación, al eliminar el "ruido documental" derivado de la utilización de los Sistemas de Clasificaciones tradicionales como medio de acceso en las recuperaciones documentales.

Apoyándonos en la difusión de las tesis de Ranganathan sobre la disolución de las diferencias entre los Lenguajes Controlados Documentales, opinamos que los Sistemas de Clasificaciones se deberían tomar como base de todos ellos para superar su multiplicidad y potenciar una ordenación lógica de los Sistemas Naturales que utilice una Clasificación Semántica mediante facetas. Esta hipótesis es apoyada desde 1955 por el CRG (Clasification Research Group), reconocida por Vickery[24] en su proposición de un Sistema de Clasificación para la Ciencia y Tecnología, y retomada por Beghtol[25] como "Conocimiento Público Interdisciplinar".

Sin embargo, dado que la ordenación de las Ciencias Humanas es bastante más compleja que la de los Sistemas Naturales, se debería profundizar en la aproximación dada por la Teoría Sistémica para el establecimiento de relaciones entre categorías, como método de entendimiento analítico (tesauros) y sintético (clasificaciones) con una aproximación jerárquica que se mueva de lo más particular (clasificación de objetos) a lo más global (clasificación de temas) y viceversa.

Siguiendo al precursor de la Teoría General de Sistemas Emery Ackoff[26], dado que "cualquier sistema puede estudiarse en términos de partes y elementos", podríamos hacer coincidir las partes y la jerarquía de estructuras

de estos sistemas con los postulados de cualquiera de los Sistemas Clasificatorios Facetados más universalmente reconocidos.

3.2. Reutilización de los Sistemas de Clasificaciones en la compilación de Tesauros.

La necesidad de establecer una normalización de los diferentes Lenguajes Controlados: sistemas clasificatorios, listas de encabezamientos, tesauros, etc., es ampliamente reconocida a escala nacional e internacional[27]. Los intentos de establecer un eje conceptual integrador han desarrollado unas terminologías básicas, agrupadas en los Macrotesauros de las diferentes Organizaciones Internacionales (ej. OCDE, OIT, IRRD, etc.), que posibilitan su enlace con los Tesauros Especializados.

Sin embargo, el no partir de una clasificación conceptual ampliamente consensuada, basada en:

- Algunos de los grandes Sistemas de Clasificaciones, desarrollados a partir de unos Principios Teóricos bien fundamentados (ej.: la clasificación de Bliss (1933), la de Sayers (1962), la Clasificación Colonada de Ranganathan en su enunciación dinámica (1967), etc.), y en
- Una codificación numérica normalizada (ej.: la Clasificación Decimal de Dewey o la Clasificación Decimal Universal), en la que poder insertar las categorías de los tesauros para su tratamiento uniforme;

Ha conducido a que *los esfuerzos globalizadores hayan sido poco efectivos,* y proliferen los Lenguajes Controlados dependientes de diferentes ordenaciones lógicas, ideológicas o del contexto sociocultural.

Aunque la deseada universalidad sea más bien una aspiración de la cultura occidental que no responde a una Clasificación Científica del Conocimiento desde la óptica de los países orientales, sería de desear que estas clasificaciones universalmente difundidas fueran utilizadas, como materia inicial, para la primera fase de la construcción de los Tesauros Documentales. Siguiendo esta metodología, Aitchison utilizó la Clasificación Bibliográfica de Bliss como fuente para la derivación de un tesauro para el Departamento de Salud y Seguridad Social de UK, en 1986[28].

Para Scibor[29], la posibilidad de utilizar la CDU como un lenguaje "intermedio" para la conexión de los tesauros especializados con otros Lenguajes Controlados, sería la razón principal de la elaboración de tablas de concordancia entre la CDU y los tesauros, para el empleo de éstas como una especie de diccionario multilingüe[30].

Las últimas tendencias detectan una menor diferenciación entre ambos tipos de Lenguajes Controlados, por ej.: en el caso del MeSH (Medical Subject Headings)[31] son las Listas de Encabezamientos de Materia las que se han convertido en Tesauros, mientras que en el caso del BSI Root Thesaurus[32] es éste el que ha dado su origen a un Sistema de Clasificación.

Con esta reutilización bidireccional, se tiende al desarrollo de mejores herramientas terminológicas para la indización y recuperación integrada en los Sistemas Integrados de Procesamiento y Recuperación de la Información, y a que los Sistemas de Clasificaciones salgan de su confinamiento en las grandes

bibliotecas y se integren con los más recientes Tesauros Conceptuales diseñados principalmente para la recuperación de la información.

3.3. Disolución de las diferencias entre los diferentes Lenguajes Documentales.

Se ha consolidado la tendencia a la disolución de las diferencias entre los diversos Lenguajes Documentales, apoyada en la difusión de los Sistemas de Clasificaciones como elementos de partida para la compilación de tesauros especializados.

Los tesauros tradicionales deben ser redefinidos en un nuevo tipo de Tesauros Conceptuales, propuestos por autores como Bates[33], Schmitz-Esser[34] o Milstead[35], que faciliten la Recuperación de la Información. Éstos, se caracterizarán por ser tesauros postcoordinados, compilados a partir del sublenguaje científico de los documentos, y dirigidos a entender el Lenguaje Natural del usuario. Para ello, deberán incorporar nuevos tipos de relaciones asociativas dependientes del documento, típicas del Lenguaje Natural, que les permita adaptarse a su nuevo tipo de usuarios, los Sistemas Expertos.

En estos Tesauros Conceptuales se deberá incluir un número de términos no controlados proporcional al de los descriptores, incluso los términos coloquiales, e intensificar las relaciones asociativas entre ambos tipos de conceptos.

En ellos, las relaciones de equivalencia serán muy importantes, dado que cuantos más sinónimos contenga un tesauro más se tomarán en consideración las distintas formas de designar un concepto, y se convertirá en más eficaz para su utilización en la recuperación por grupos de usuarios diferenciados.

Las relaciones asociativas sintácticas (propias del contexto documental) son más importantes que las de equivalencia, dado que la lógica booleana de recuperación será sustituida por los modelos conceptuales que utilizan el Lenguaje Natural (obtenido del Análisis Contextual de los documentos) en la enunciación de las ecuaciones de búsqueda[36].

Se partiría de las bases de datos documentales y mediante análisis semántico (indización semántica latente, etc.[37]) se irían extrayendo conceptos, siguiendo un criterio relacional previamente definido. A partir de estos conceptos se desarrollarían diferentes Tesauros Conceptuales, unidos en una red semántica de estructuras neuronales en la que cada nodo contenga una serie de descriptores asociados con un único concepto semántico que pueda ser igualmente identificado en la red hipertextual de “pequeñas piezas de información” textual del espacio documental.

4. ONTOLOGÍAS PARA LA INTEGRACIÓN TERMINOLÓGICA DEL ESPACIO WEB.

Tom Gruber define una ontología como:

“Una especificación explícita de un concepto...Por razones prácticas, decidimos escribir una ontología como una serie de definiciones de vocabulario formalizado (distinto del concepto filosófico tradicional)”.

> "Para un sistema de IA (inteligencia artificial), lo que "existe" es lo que se puede representar. Cuando el conocimiento de un dominio científico se expresa en forma declarativa, la serie de objetos que pueden representarse se denomina el universo del discurso. La serie de objetos y las relaciones que se pueden describir entre ellos se reflejan en el vocabulario con el que se representa el conocimiento"[38].

De su definición se deduce una cierta analogía con otras estructuras conceptuales de representación de la información del tipo de: los tesauros o las clasificaciones conceptuales, que igualmente establecen asociaciones entre sus conceptos. Sin embargo, existen diferencias fundamentales en cuanto a *la representación del conocimiento:*

a) Las ontologías permiten representar axiomas (conocimientos ciertos inmutables) y razonar con ellos mediante reglas de inferencia, definidas en "el núcleo del sistema" de los Agentes Expertos que filtran la información[39].

b) Además, pueden ser reutilizadas para la autogeneración de tesauros conceptuales internos que permitan distinguir sinónimos, suprimir homónimos e inducir relaciones asociativas entre los descriptores.

Para mejorar la precisión de las recuperaciones documentales, se propone el diseño de ontologías especializadas (a partir de las bases de datos con texto completo de las diversas áreas del conocimiento) que enlacen con las preguntas específicas del área tratada. Una ontología para una base de conocimientos de la IA abarcará los diferentes tipos de documentos, descripciones conceptuales, sus relaciones asociativas, los diferentes problemas científicos que plantean; además de índices, descripciones bibliográficas, tesauros, códigos clasificatorios, formalizaciones de validez, información terminológica, etc. Su aplicación debe proporcionar una metavisión de la estructura y de la terminología del dominio que facilite recuperaciones altamente relevantes.

Para garantizar la universalidad, consistencia y concisión del conocimiento en los procesos de desarrollo de los sistemas de IA, se debe diferenciar claramente el diseño de una ontología del de las bases de conocimientos en las que se integren, a pesar de que ambas contengan "conocimiento":

a) Pues, en cuanto al propósito de *la codificación de su conocimiento,* las ontologías deben diseñarse con suficiente abstracción y generalidad para compartir y reutilizar su conocimiento.

b) Y en cuanto *la especificación de los requerimientos,* las ontologías deben llevar una especificación inicial de su alcance para un dominio dado que les permita su reutilización independientemente del comportamiento y del dominio de la aplicación que la utilice[40].

Dado que las fuentes de información en Internet son muy diversas, cualquier *sistema virtual de ontologías entre sedes Web* debería ser muy general y polivalente en el primer nivel de la jerarquía. Se integraría una superontología dinámica (o catálogo de catálogos), en continua evolución a partir de otras subontologías que se adapten y sobrevivan en sus propias áreas de trabajo habitual; y se utilizarían aplicaciones informáticas de la IA que permitieran la automatización del proceso de autogeneración de tesauros conceptuales de abajo-arriba.

CONCLUSIONES Y APUESTAS DE FUTURO:

El crecimiento acelerado de la Web ha producido una explosión de la variedad de fuentes disponibles. Los profesionales de la información han revalorizado las técnicas de los Sistemas de Clasificaciones Conceptuales para el desarrollo de catálogos digitales (ej.: del tipo Yahoo, Excite, etc.), que organicen con precisión la información descentralizada del espacio documental.

1) Se necesita un modelo integrado de procesamiento de la información que parta de la "interacción total" entre los documentos y las preguntas, los considere como entidades del mismo tipo, y normalice sus conceptos. Para generar un Espacio Conceptual Integrado de Procesamiento de la Información, que refleje las relaciones asociativas entre conceptos más estrechas e ignore las menos relevantes, se debe ponderar el grado de solapamiento entre conceptos ("clustering matching process") y extraer los que incrementen la relevancia de cada recuperación.

Para el dilatado proceso de la Gestión del Conocimiento, se propone la adopción del Modelo Conceptual que, mediante el razonamiento inductivo-inferencial, genere en los Agentes Expertos de la IA sus propias reglas de adaptación a las preguntas no explícitas "a priori" por el usuario, y sustituya al Método Deductivo basado en reglas jerárquicas.

El desarrollo de un Modelo Conceptual debe incluir potentes Estructuras de Representación del Conocimiento, capaces de integrar: a) todo tipo de relaciones, b) sus participantes y c) sus actuaciones, que faciliten la búsqueda de la información documental por su contenido (conocimiento), a partir de la identificación de sus componentes contextuales y de su almacenamiento en los Esquemas de Representación de la Base de Conocimientos[41].

2) En estos momentos se vuelve a pensar en los tesauros como herramienta de precisión para la recuperación del conocimiento que circula por la red Internet.

Estos últimos tesauros postcoordinados difieren de los tesauros precoordinados preferidos en las décadas de los sesenta y setenta en que: su compilación, se hace a partir del sublenguaje científico de contextos concretos, y en que su utilización es preferentemente para la recuperación de documentos.

Estos nuevos tesauros son un tipo de macroserie del sublenguaje controlado en un dominio científico específico, que se usan durante: el proceso de indización, como ayuda en la identificación de los conceptos, y en el proceso de recuperación, como fuente de nuevos términos que identifiquen conceptos y aumenten la precisión de las búsquedas.

Se integran en los interfaces de los Sistemas de Procesamiento de la Información para mejorar la pertinencia de las búsquedas, debido a las numerosas relaciones asociativas y contextuales que presentan. Estos sistemas hacen uso de los Tesauros Conceptuales como elemento de precisión del lenguaje mixto usado en sus documentos.

Para su autogeneración se deben utilizar potentes algoritmos de IA, capaces de procesar contextualmente el Lenguaje Natural y extraer los

conceptos normalizados de los documentos de cada Base de Datos Documental.

3) Como hemos comentado al final del apartado segundo, la dicotomía entre Lenguajes Controlados y Lenguaje Natural está dando paso a la integración del lenguaje del usuario con los lenguajes documentales. Existe, además, una clara tendencia a la conjunción de los lenguajes documentales con la terminología, ya que los textos técnicos se almacenan en soporte digital y pueden convertirse a un formato adecuado para el análisis terminológico de córporas especializados.

Estos Lenguajes Controlados parten de las Clasificaciones Conceptuales (ej.: ahora Yahoo) para autogenerar los tesauros hipertextuales de Internet, mediante la extracción de la terminología de los documentos a texto completo y su integración en las ontologías terminológicas de dicha red.

A partir de esta metodología, diferentes profesionales de la información (como Chan[42], Vizine-Goetz[43] o McKiernan[44]) han señalado las ventajas de utilizar ontologías reutilizables para organizar el universo de Internet. Encadenándolas con los Hipertesauros Conceptuales en un espacio terminológico de redes neuronales, en el que investigadores como Kohonen o Xia Lin[45] han empezado a desvelar sus analogías con el conocimiento del mundo aportado por las neuronas humanas.

La finalidad sería potenciar los Tesauros Jerárquicos, existentes desde hace décadas en cualquier subárea del conocimiento, y convertirlos en Tesauros Conceptuales que contengan numerosas relaciones asociativas entre descriptores y relaciones de equivalencia entre descriptores y no descriptores. Tras su posterior conversión en tesauros hipertextuales, mediante el lenguaje HTML o su sucesor el XML, podremos conectarlos entre sí (en la Web de Internet) de manera que formen una poderosa y reutilizable ontología terminológica, representada en el espacio conceptual hipertextual como una red semántica neuronal[46].

Esta ontología de Lenguajes Controlados Hipertextuales permitirá enlazar de forma descentralizada todo tipo de objetos en el espacio distribuido de la información contenida en la Web, para su recuperación con los conceptos establecidos en las ecuaciones de búsqueda. Esta información desplegará sus elementos textuales en forma de asociación de redes documentales, a partir de una indización dinámica de abajo-arriba que le permita la confrontación de sus conceptos con la red terminológica anterior[47].

BIBLIOGRAFÍA:

[1] SAUSURRE, F. de. Curso de Lingüística General (Trad. del francés: París: Fayot, 1916). Barcelona: Planeta-Agostini, 1985, p. 51.

[2] PINTO, Mª; GÁLVEZ, C. "Hacia una teoría integradora de la clasificación documental". En: Manual de Clasificación Documental. Madrid: Síntesis, 1987, p. 39.

[3] CHOMSKY, N. Transformational Analysis. Tesis doctoral, Universidad de Pennsylvania, 1955.

[4] VAN DIJK, T. A. Texto y Contexto: semántica y pragmática del discurso. Madrid: Cátedra, 1988.

[5] MANIEZ, J. Los lenguajes documentales y de clasificación: Concepción, construcción y utilización en los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, 1992.

[6] ROWLEY, J. E. Organizing Knowledge: an introduction to information retrieval. Aldershot: Ashgate, 1992 (2ª ed. rev.), pp. 176-177.
[7] ESTEBAN NAVARRO, M. A. "Fundamentos epistemológicos de la Clasificación Documental", SCIRE, 1995, 1 (1), pp. 90-91.
[8] ZSILKA, T. "Communicative relationships of the text", Text vs sentence, 1981 (3), pp. 231-240.
[9] SHREIDER, Y. A.; USPENSKY, V. A. "Semantic aspects of informatics". En: Research on the theoretical basis of information, Moscow: FID, 1975, pp. 152-169.
[10] SVENONIUS, E. "Ranganathan and Classification Science". Libri, 1992, 42 (3), pp. 176-183.
[11] NEUMANN, R. "Thesauri und Klassifikation von Wissen". En: Das Lexicon in der Grammatik-die Grammatik im Lexikon, Hamburg: Helmut Buske, 1977, pp. 163-277.
[12] GREEN, R. "Topical Relevance Relationships. Why Topic Matching Fails?". Journal of the ASIS, 1995, 46 (9), p. 648.
[13] MANIEZ, J. Le rôle de la syntaxe dans les systèmes de recherche documentaire. IUT de Dijon, Depártement Carrières de l'information, 1976, 2 vols., 184+182 pp.
[14] FUGMANN, R. "The Complementarity of Natural and Indexing Languages". International Classification, 1982 (9), p. 142.
[15] DUBOIS, C. P. R. "Free text vs controlled vocabulary: a reassessment". ON LINE Review, 1987 (11), p. 248.
[16] SVENONIUS, E. "Unanswered Questions in the Design of Controlled Vocabularies". Journal of the ASIS, 37(5), 1986, pp. 331-340.
[17] KRISTENSEN, J. "Expanded end-user's query statements for free text searching with a search-aid thesaurus". Information Processing & Management, 1993, 29 (6), pp. 733-744.
[18] FREMER, E.; LARSSON, B. "SPIRS, WinSPIRS, and OVID: a question of free-text versus thesaurus retrieval?" [carta]. Bull. Med. Assoc., 1997, 85(1), pp. 57-58.
[19] CROFT, W. B. y DAS, R. "Experiments with query acquisition and use in document retrieval systems". En: Proceeedings 13th Conference on Research and Development in Information Retrieval, Brussels, Belgium, Sept.1990.
[20] EKMEKCIOGLU, F. C., ROBERTSON, A. M. Y WILLET, P. "Effectiveness of query expansion in ranked-output document retrieval systems". Journal of Information Science, 1992, 18, pp.139-147.
[21] VOORBIJ, H.J. "Title keywords and subject descriptors: a comparison of subject search entries of books in the humanities and social sciences". Journal of Documentation, 1998, 54 (4), pp. 466-476.
[22] VIZINE-GOETZ, D.; MITCHELL, J.S. "Dewey 2000. Cataloging Productivity Tools". En: Annual Review of OCLC Research, Dublin: OCLC, 1997.
[23] MOREIRO GONZÁLEZ, J.A. "De la Documentación a la Ciencia de la Información: evolución de los conceptos y aplicaciones documentales". Seminario de Humanidades Agustín Millares Carlo, separata Homenaje a Antonio de Bethencourt Massieu, 1995, cita nº 37 y p. 20.
[24] VICKERY, B.C. "Knowledge Representation: a brief review". Journal of Documentation, 1986, 42 (3), pp. 145-159.
[25] BEGHTOL, C. "Facets as interdisciplinary undiscovered public knowledge". Journal of Documentation, 1995, 51 (3), pp. 194-224.
[26] ACKOFF, E. On purposeful systems. Aldine-Atherton, Chicago, 1972.
[27] IZQUIERDO ARROYO, J.M.; MORENO FERNANDEZ, L.M. "Listas de encabezamientos de materia y Thesauri en perspectiva comparada". Documentación de las Ciencias de la Información, 1994, 17, pp. 287-310.
[28] AITCHISON, J. "A classification as a source for a thesaurus: the bibliographic classification of H.E. Bliss as a source of thesaurus terms and structure". Journal of Documentation, 1986, 42 (3), pp. 160-181.
[29] SCIBOR, E. "La CDU y los Thesauri: Diferentes aspectos del problema". Boletín de la ANABAD, 1978, 28 (2), pp. 81-92.
[30] MORENO FERNÁNDEZ, L.M. "Una vez más: La CDU no es un Thesaurus". Documentación de las Ciencias de la Información, 1992, 15, pp. 67-81.
[31] National Library of Medicine. Medical subject headings. Washington DC: Government Printing Office, 1972.
[32] BRITISH STANDARDS INSTITUTION. BSI ROOT thesaurus. Milton Keynes: BSI, 1981.

[33] BATES, M. J. "Subject Access in Online Catalogs: A Design Model". Journal of ASIS, 1986, 37 (6), p. 361.
[34] SCHMITZ-ESSER, W. "New Approaches in Thesaurus Application". International Classification, 1991, 18 (3), pp. 143-147.
[35] MILSTEAD, J.L. “Invisible Thesauri: the year 2000”. ONLINE & CDROM Review, 1995, 19 (2), pp. 93-94.
[36] GREEN, R. “The Role of Relational Structures in Indexing for the Humanities”. Knowledge Organization, 1997. 24(2), pp. 72-83.
[37] CAID, W.R. et al. “Learned vector-space models for document retrieval”. Information Processing &Management, 1995, 31(3), pp. 419-429.
[38] GRUBER, T.R. “A Translation Approach to Portable Ontologies”. Knowledge Acquisition, 1993, 5(2), pp. 199-220.
[39] LÓPEZ ALONSO, M.A. “Legal Databases with Hypertext Organization Systems”. En: Quinto Convegno Internazionalle dell’Instituto para la Documentazione Giuridica, Firenze: CNR, dec. 1998.
[40] GUERRERO BOTE, V.; LOZANO TELLO, A. “Vínculos entre Ontologías y la Biblioteconomía y Documentación”. En: Actas del IV Congreso ISKO-España, Granada, 1999, pp. 25-31.
[41] CRAVEN, M. et al. “Learning to Extract Symbolic Knowledge from the World Wide Web”. En: Proceedings 15th National Conference on Artificial Intelligence, Madison: WI, 1998, AAAI Press.
[42] CHAN, L.M. “Classification, present and future”. Cataloging & Classification Quarterly, 1995, 21 (2), .5-17.
[43] VIZINE-GOETZ, D. “Using library classification schemes for Internet resources”. Proceedings of the OCLC Internet Cataloging Colloquium. Dublin, Ohio: OCLC, 1996.
[44] McKIERNAN, G. “The new/old World Wide Web order: the application of “neo-conventional” functionality to facilite access and use of a WWW database of science and technology Internet resources”. Journal of Internet Cataloging, 1997, 1 (1), pp. 47-55.
[45] ZAVREL, J. “Neural Navigation Interfaces for Information Retrieval: Are They More Than an Appealing Idea?”. Artificial Intelligence Review, 1996, 10 (5-6), pp. 477-504.
[46] Plumb Design Inc. Plumb Design Visual Thesaurus (Thinkmap™). http://www.thinkmap.com (visitado: 27/02/2001).
[47] Van DOORN, M.G.L.M. Thesauri and the Mirror retrieval model. Holanda: University of Twenty, Database Group. Master Thesis, Jul. 1999.